



TESIS - TE142599

***SENTIMENT ANALYSIS MENGGUNAKAN
SUPPORT VECTOR MACHINE(SVM)***

PETRIX NOMLENI
NRP 2213206717

DOSEN PEMBIMBING
Mochamad Hariadi, ST., M.Sc., Ph.D
Dr. I Ketut Eddy Purnama, ST., MT

PROGRAM PASCA SARJANA
BIDANG KEAHLIAN TELEMATIKA (CIO)
JURUSAN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2015



TESIS-TE142599
SENTIMENT ANALYSIS USING
SUPPORT VECTOR MACHINE(SVM)

PETRIX NOMLENI
NRP 2213206717

SUPERVISOR
Mochamad Hariadi, ST., M.Sc., Ph.D
Dr. I Ketut Eddy Purnama, ST., MT

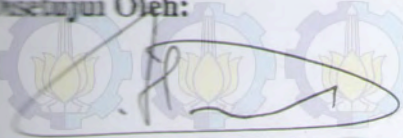
MAGISTER PROGRAM
TELEMATIC ENGINEERING
ELECTRICAL ENGINEERING DEPARTMENT
FACULTY OF INDUSTRIAL TECHNOLOGY
SEPULUH NOPEMBER INSTITUTE OF TECHNOLOGY
SURABAYA
2015

**Magister Teknik (MT)
di
Institut Teknologi Sepuluh Nopember**


**Oleh :
Petrix Nomleni
2213 206 717**

**Tanggal Ujian : 16 Januari 2015
Periode Wisuda : Maret 2015**


Disetujui Oleh:


1. Mochamad Hariadi, ST., M.Sc., Ph.D
NIP: 196912091997031002


(Pembimbing I)


2. Dr. I Ketut Eddy Purnama, ST., MT
NIP: 196907301995121001


(Pembimbing II)


3. Prof. Dr. Ir. Mauridhi Hery Purnomo, M.Eng.
NIP: 195809161986011001

(Penguji)



4. Dr. Ir. Yoyon Kusnendar Suprpto, M.Sc.
NIP: 195409251978031001

(Penguji)


5. Dr. Eko Mulyanto Yuniarno, ST., MT.
NIP: 196806011995121009

(Penguji)

Direktur Program Pascasarjana


Prof. Dr. Ir. Adi Soeprijanto, MT
NIP: 196404051990021001

SENTIMENT ANALYSIS MENGGUNAKAN SUPPORT VECTOR MACHINE(SVM)

Nama Mahasiswa : Petrix Nomleni
NRP : 2213206717
Pembimbing I : Mochamad Hariadi, ST., M.Sc., Ph.D
Pembimbing II : Dr. I Ketut Eddy Purnama, ST., MT

ABSTRAK

Pemerintah sebagai pelayan masyarakat memiliki peran yang sangat besar dalam meningkatkan kesejahteraan masyarakat. Maka perlu diadakan suatu perbaikan secara bertahap guna meningkatkan pelayanan masyarakat (*public services*) sebagai tugas utama pemerintah, untuk itu perlu adanya sikap keterbukaan dari pemerintah untuk dapat menerima setiap keluhan masyarakat mengenai kebijakan / program yang langsung menyentuh kepentingan masyarakat.

Media Center merupakan sistem pelayanan informasi yang terintegrasi kepada masyarakat untuk ikut berpartisipasi dalam pembangunan dengan berbagai cara seperti ide, pengaduan, keluhan, kritik, saran dan pertanyaan. Untuk itu perlu adanya klasifikasi untuk sentiment analysis keluhan masyarakat informasi yang masuk ke media center sehingga pengelola dapat memberikan informasi yang efisien dan tepat kepada masyarakat dan pemerintah dapat mengetahui bidang mana yang perlu dibenahi dalam pembangunan.

Sentiment analysis merupakan proses klasifikasi dokumen tekstual ke dalam beberapa kelas seperti sentimen positif dan negatif serta besarnya pengaruh dan manfaat dari sentiment analysis tersebut. Pada penelitian ini dibahas klasifikasi keluhan masyarakat terhadap pemerintah pada media sosial *facebook* dan *twitter* sapawarga data berbahasa Indonesia menggunakan metode *Support Vector Machine* (SVM) yang dijalankan dalam komputasi terdistribusi dengan menggunakan *Hadoop*. Pengujian dilakukan dengan perhitungan *precision*, *recall*, *F-Measure* serta akurasi dengan menghasilkan akurasi rata-rata diatas 80% dengan akurasi tertinggi 84.4086% *precision* 81% *recall* 84% serta *F-Measure* 80%.

Kata kunci: *Media Center*, *Support Vector Machine*, klasifikasi, *sentiment analysis*

SENTIMENT ANALYSIS USING SUPPORT VECTOR MACHINE(SVM)

Name : Petrix Nomleni
NRP : 2213206717
Supervisor : Mochamad Hariadi, ST., M.Sc., Ph.D
Co-supervisor : Dr. I Ketut Eddy Purnama, ST., MT

ABSTRACT

Government as a public servant has a very big role in improving the welfare of society. So there should be a gradual improvement in order to improve public services as the main task of government, to the need for openness of government to be able to receive any complaints about the policies / programs that directly touch the interests of the community.

Media Center is a system of integrated information services to the public to participate in the development of a variety of ways such as ideas, complaints, complaints, criticisms, suggestions and questions. For that we need a classification for sentiment analysis complaints that information into the media center so that managers can provide an efficient and precise information to the public and the government can determine what areas need to be addressed in development.

Sentiment analysis is the process of classification of textual documents into several classes such as positive and negative sentiment as well as the magnitude of the effect and the benefits of sentiment analysis. In this study discussed the classification of public complaints against the government on facebook and twitter social media sapawarga Indonesian language of data using Support Vector Machine (SVM) which is executed in a distributed computing using Hadoop. Testing is done with the calculation precision, recall, F-Measure and accuracy with average accuracy above 80 % with the highest accuracy 84.4086 % precision 81 % recall of 84 % and F-Measure 80 %.

Keywords: Media Center, Support Vector Machine, classification, sentiment analysis

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yesus Kristus atas anugerah dan berkat-Nya penulis dapat menyelesaikan Tesis yang berjudul : "SENTIMENT ANALYSIS MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM)" ini disusun sebagai salah satu syarat untuk menyelesaikan pendidikan S-2 Bidang Keahlian Telematika (CIO) Program Pascasarjana Teknik Elektro, Institut Teknologi Sepuluh Nopember , Surabaya.

Selesaiannya Tesis ini tidak lepas dari bantuan banyak pihak, baik secara langsung maupun tidak langsung. Untuk itu penulis mengucapkan banyak terimakasih kepada :

1. Istriku tercinta Miegdal A.H. Nomleni, yang telah memberi izin untuk melanjutkan S2 dan selalu membantu serta memberikan dukungan baik dorongan moril maupun spiritual dalam menyelesaikan Tesis ini.
2. (Almh) ibunda tercinta yang semasa hidupnya selalu memberikan semangat dan doanya kepada penulis.
3. Bapak dan ibu mertua penulis, dan saudara-saudara, yang selalu memberikan doa, dukungan dan semangat yang tak pernah ada habisnya.
4. Bapak Mochamad Hariadi, ST., M.Sc., Ph.D., dan Bapak Dr. I Ketut Eddy Purnama, ST., MT., selaku dosen pembimbing yang selalu memberikan arahan selama pengerjaan Tesis.
5. Bapak-bapak dosen penguji, yakni bapak Prof. Dr. Ir. Mauridhi Hery Purnomo, M.Eng., Bapak Dr. Ir. Yoyon Kusnendar Suprpto, MSc., dan Bapak Dr. Eko Mulyanto Yuniarno, ST., MT., atas masukan yang diberikan sehingga tesis ini dapat menjadi lebih baik.
6. Keluarga besar S2 Telematika CIO 2013, yang dari awal perkuliahan saling memberikan semangat dan saling mendukung dalam setiap aktivitas perkuliahan.
7. Kantor Pengolahan Data Elektronik (KPDE) Provinsi Nusa Tenggara Timur, yang mendukung moril dan spiritual melalui data dan informasi lisan maupun tulisan.
8. Kementerian Komunikasi dan Informatika, yang memberikan beasiswa sehingga penulis, sehingga membantu penulis dalam hal pendanaan selama masa perkuliahan.
9. Semua pihak yang telah membantu penulis dalam menyelesaikan Tesis ini.

Penulis menyadari bahwa tulisan ilmiah yang tertuang dalam Tesis ini masih jauh dari kesempurnaan baik dari bentuk penyusunan maupun materinya. Kritik dan saran yang membangun sangat penulis harapkan untuk kemajuan ilmu pengetahuan dan teknologi. Penulis berharap, semoga tulisan ilmiah yang tertuang dalam buku Tesis ini, dapat memberikan informasi dan manfaat bagi peminat peneliti selanjutnya, pembaca pada umumnya dan mahasiswa Jurusan Teknik Elektro pada khususnya.

Surabaya, Januari 2015
Penulis

Daftar Isi

Lembar Pengesahan	iii
Pernyataan	v
Abstrak	vii
Abstract	ix
Kata Pengantar	xi
Daftar Isi	xiii
Daftar Gambar	xvii
Daftar Tabel	xix
Daftar Istilah	xxi
1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Permasalahan	2
1.3 Batasan Masalah	2
1.4 Tujuan Dan Manfaat Penelitian	3
1.4.1 Tujuan	3
1.4.2 Manfaat	3
1.5 Sistematika Penulisan	3
2 KAJIAN PUSTAKA DAN LANDASAN TEORI	5
2.1 <i>Sentiment Analysis</i>	5
2.2 <i>Text Mining</i>	5
2.3 Riset Terkait	6
2.4 <i>Media Center</i>	7
2.5 <i>Text Preprocessing</i>	8
2.5.1 Pembersihan Dokumen(<i>Clansing</i>)	8
2.5.2 <i>Parsing</i>	8
2.5.3 <i>Tokenizing</i>	8
2.5.4 <i>Stopword Removal</i>	9
2.5.5 <i>Stemming</i>	9

2.5.6	Normalisasi	9
2.6	<i>Big Data</i>	9
2.7	<i>Hadoop</i>	10
2.7.1	HDFS (<i>Hadoop Distributed File System</i>)	11
2.7.2	<i>MapReduce</i>	12
2.8	Klasifikasi Data	13
2.8.1	Model Klasifikasi	13
2.8.2	Tujuan Klasifikasi	13
2.8.3	Tahapan Proses klasifikasi data	14
2.9	<i>Algoritma Support Vector Machine</i>	16
2.9.1	Klasifikasi <i>Linear</i>	16
2.9.2	Klasifikasi <i>Non Linear SVM</i>	19
2.10	Precision, Recall, F-measure	22
3	METODOLOGI PENELITIAN	25
3.1	Metodologi Penelitian	25
3.2	Perancangan Sistem	26
3.3	<i>Diagram Alir Preprocessing</i>	27
3.4	<i>Diagram Alir Filtering/Stopword Removal</i>	28
3.5	Pembobotan	28
3.5.1	Term Weighting	28
3.5.2	Term Presense (TP)	29
3.5.3	Inverse Document Frequency (IDF)	29
3.5.4	TF-IDF	29
3.6	Proses Klasifikasi (SVM)	30
3.7	<i>Precision, Recall dan F-Measure</i>	37
4	HASIL DAN PEMBAHASAN	39
4.1	Proses Akuisisi Data	39
4.2	Preprocessing	40
4.2.1	Proses Pembersihan Dokumen (<i>Cleansing</i>)	40
4.2.2	Proses <i>Case Folding</i>	40
4.2.3	Proses <i>Parsing</i>	41
4.2.4	Proses <i>Filtering/Stopword Removal</i>	42
4.2.5	Proses <i>Stemming</i>	43
4.2.6	Proses Pembobotan	43
4.3	Klasifikasi dengan <i>Support Vector machine (SVM)</i>	45

4.3.1	Percobaan Dengan Data Pelatihan 80% dan Data Pengujian 20%	48
4.3.2	Percobaan Dengan Data Pelatihan 70% dan Data Pengujian 30%	49
4.3.3	Percobaan Dengan Data Pelatihan 60% dan Data Pengujian 40%	51
4.3.4	Percobaan Dengan Data Pelatihan 50% dan Data Pengujian 50%	53
4.3.5	Percobaan Dengan Data Pelatihan 40% dan Data Pengujian 60%	54
4.3.6	Percobaan Dengan Data Pelatihan 30% dan Data Pengujian 70%	56
4.3.7	Percobaan Dengan Data Pelatihan 20% dan Data Pengujian 80%	57
5	KESIMPULAN DAN SARAN	61
5.1	Kesimpulan	61
5.2	Saran	61

Daftar Pustaka	63
-----------------------	-----------

Daftar Tabel

2.1	Tabel Kontingensi	23
3.1	Tabel Data Set	36
3.2	Tabel Confusion Matrix	37
4.1	Akuisisi Data	39
4.2	Pembersihan Dokumen	40
4.3	Proses Case Folding	41
4.4	Parsing	41
4.5	Filtering	42
4.6	<i>Filtering 1</i>	43
4.7	<i>Stemming</i>	43
4.8	Proses Pembobotan	44
4.9	Split Data Pelatihan dan Data Pengujian	45
4.10	<i>Confusion Matrix</i>	46
4.11	Hasil Percobaan Pertama	48
4.12	Hasil Percobaan Kedua	49
4.13	Hasil Percobaan Ketiga	51
4.14	Hasil Percobaan Keempat	53
4.15	Hasil Percobaan Kelima	54
4.16	Hasil Percobaan Keenam	56
4.17	Hasil Percobaan Ketujuh	57

Daftar Pustaka

- Bing Liu. 2010. Sentiment Analysis and Subjectivity, in Handbook of Natural Language Processing
- Edwin Lunando, Ayu Purwarianti, 2013, *Indonesian Social Media Sentiment Analysis with Sarcasm Detection*,
- Imam Fahrur Rozi, Sholeh Hadi Pramono, Erfan Achmad Dahlan, 2012, Implementasi Opinion Mining (Analisis Sentimen) Implementasi untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi
- Judith Hurwits, Alan Nugent, Dr. Fern Helper, Marcia Kaufman, 2013, *Big Data For Dummies A Wiley Brand*
- Krisantus, 2007, Tutorial SVM
- Muhamad Yusuf Nur dan Diaz D. Santika, 2011, *Analisa Sentimen Pada Dokumen Bahasa Indonesia Dengan Pendekatan (Support Vector Machine)*,
- Ni Wayan Suhartini Saraswati, I. Ketut Gede Darma Putra, Ni Made Ary Esta Dewi Wirastuti, 2011, Text Mining Dengan Metode Naive Bayes Classifier dan Support Vector Machine Untuk Sentiment Analysis
- Noviah Dwi Putranti dan Edi Winarko, 2013, *Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine*,
- Prima Arfinada Putri, Achmad Ridok, Indriati, 2013, *Implementasi Metode Improved K-Nearest Neighbor Pada Analisis Sentimen Twitter Berbahasa Indonesia*
- Romen Feldman, James Sanger, 2007, *The Text Mining Handbook (Advanced Approaches In Analyzing Unstructured Data)*
- Sang-Hyun Cho, Hang-Bong Kang *Statistical Text Analysis and Sentiment Classification in Social Media*
- Santosa Budi *Tutorial Support Vector Machine*
- <http://yonathan.web.ugm.ac.id/stem/?stem>. Diakses pada tanggal 15 Desember 2014

Daftar Gambar

2.1	SOP Media Center	7
2.2	<i>Hadoop distributed File System</i>	10
2.3	<i>MapReduce</i>	12
2.4	Model Klasifikasi	13
2.5	Proses Klasifikasi	15
2.6	<i>Algoritma Support Vector Machine</i>	16
2.7	Klasifikasi <i>Linear SVM</i>	17
2.8	Klasifikasi <i>Linear SVM</i>	17
2.9	Klasifikasi <i>Non Linear SVM</i>	20
2.10	Transformasi dari vektor input ke <i>feature space</i>	21
3.1	Rancangan Sistem Klasifikasi	26
3.2	Diagram Alir <i>Diagram Alir Preprocessing</i>	27
3.3	Diagram Alir <i>Diagram Alir Preprocessing</i>	28
3.4	Diagram Alir Klasifikasi dengan SVM	30
4.1	Percobaan 1	49
4.2	Percobaan 2	50
4.3	Percobaan 3	52
4.4	Percobaan 4	54
4.5	Percobaan 5	55
4.6	Percobaan 6	57
4.7	Percobaan 7	58

BIOGRAFI PENULIS



Nama : Petrix Nomleni
TTL : Kupang, 10 Maret 1976
Agama : Kristen Protestan
Alamat I : Jl. TDM II No. 1 Kupang - NTT
Alamat II : Semoluwaru Selatan Gang II No 4, Surabaya
HP : 081236221464
Email : petrix_nomleni@yahoo.com
Petrix13@mhs.ee.its.ac.id

Jenjang Pendidikan :

1. Tahun 1983 - 1989 : SD Impres Oetete 2 Kupang
2. Tahun 1989 - 1991 : SMP Negeri 1 Kupang
3. Tahun 1994 - 1997 : SMA PGRI Kupang
4. Tahun 2002 - 2007 : Universitas Janabadra - Yogyakarta
Program Studi Teknik Informatika
5. Tahun 2013- 2015 : Institut Teknologi Sepuluh November
Fakultas Teknologi Industri
Jurusan Teknik Elektro
Program Studi Telematika
Konsentrasi *Chief Information Officer*

Riwayat Pekerjaan :

1. PT. Surya Bhakti Utama (2008 – 2009)
2. Kantor Pengolahan Data Elektronik Provinsi NTT(2010 – Sekarang).

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Dalam era perkembangan teknologi informasi yang semakin pesat di Indonesia saat ini, keterbukaan atau transparansi merupakan suatu hal yang sangat penting dalam rangka melaksanakan fungsi pengontrolan. Seperti kita ketahui bahwa kehadiran pemerintah sebagai pelayan masyarakat memiliki peran yang sangat besar dalam meningkatkan kesejahteraan masyarakat. Sistem birokrasi yang ada sekarang ini yang dianggap sebagai sarang korupsi, kolusi dan nepotisme (KKN), penghambat investasi dan lain-lain. Untuk itu perlu diadakan suatu perbaikan secara bertahap guna meningkatkan pelayanan masyarakat (*public services*) sebagai tugas utama pemerintah, maka terlebih dahulu perlu adanya sikap keterbukaan dari pemerintah untuk dapat menerima setiap kritik, saran ataupun keluhan masyarakat mengenai kebijakan/program yang langsung menyentuh kepentingan masyarakat misalnya penyelewengan kebijakan dan lain-lain. Hal ini juga dianggap sangat penting untuk mengaktifkan peran masyarakat, LSM dan lain-lain sebagai suatu fungsi kontrol terhadap setiap kebijakan pemerintah.

Perkembangan teknologi yang sangat cepat, tentunya membuka peluang untuk mewujudkan harapan baru. Dengan adanya konsep *e-government* sebagai salah satu upaya yang dikembangkan untuk memperbaiki sistem birokrasi tentunya perlu dimanfaatkan semaksimal mungkin. Untuk itu pembangunan *media center* yang merupakan sistem pelayanan informasi yang terintegrasi kepada masyarakat untuk ikut berpartisipasi dalam pembangunan dengan berbagai cara seperti ide, pengaduan, keluhan, kritik, saran dan pertanyaan.

Media sosial baik itu *facebook* maupun *twitter* saat ini banyak diminati oleh setiap kalangan, yang memungkinkan penggunaannya dapat mengirimkan pesan atau menerima pesan dan dibaca secara bebas, tetapi dapat diatur untuk dilihat oleh pengguna lain yang mengikutinya atau pertemanan. Saat ini media sosial mempunyai peranan yang sangat penting dalam memberikan informasi yang sangat cepat dikarenakan pengguna media sosial semakin bertambah setiap harinya dan informasi yang dihasilkan sangat beragam seperti seperti berita, pertanyaan, opini, komentar, kritik baik yang bersifat positif maupun negatif. Kita dapat melihat bagaimana pendapat orang lain terhadap suatu permasalahan melalui sentiment analysis, sehingga dapat membantu kita dalam mengambil sebuah keputusan dengan lebih cepat dan tepat.

Media sosial juga sudah merambah ke sektor pemerintahan dimana saat ini pemerintah menggunakan media sosial untuk merespon masukan dari masyarakat terhadap pembangunan salah satunya yaitu media center yang menampung data masukan berupa informasi dari masyarakat kepada pemerintah untuk menjadi masukan perbaikan dan pembangunan agar perbaikan dan pengembangan yang dapat dilakukan dengan cepat dan berkesinambungan untuk masyarakat.

kat. Keluhan yang disampaikan oleh masyarakat merupakan masukan yang sangat bernilai dan merupakan salah satu instrumen untuk melakukan evaluasi dan deteksi dini terhadap kelemahan sistem ataupun penyimpangan pelaksanaan pembangunan.

Keluhan masyarakat yang diterima melalui *media center* memiliki tingkat tatanan bahasa dan kesopanan yang berbeda. Untuk itu perlu adanya sistem filterasi keluhan dengan cara melakukan *sentiment analysis* disetiap keluhan yang ditujukan pada Pemerintah. *Sentiment analysis* mengacu pada penerapan pengolahan bahasa alami untuk mengidentifikasi informasi secara subjektif. Sentiment analysis dilakukan untuk mengetahui sikap emosional penulis dalam kontekstual dokumen yang diklasifikasikan oleh aplikasi supaya keluhan-keluhan tersebut masuk kebagian pada kelompok sentiment yang sama.

Pengklasifikasian teks dilakukan dengan cara mengkategorikan dokumen-dokumen ke dalam satu atau beberapa dari sekumpulan topik-topik. Setelah proses pengelompokan itu selesai akan di dilakukan proses filterasi sehingga dapat ditentukan kelompok klasifikasi mana yang harus diambil dan klasifikasi mana yang harus dibuang berdasarkan tingkat kesopanan pada media center. Sebagai solusi permasalahan, maka diperlukan sebuah proses yang berjalan secara otomatis untuk melakukan perklasifikasian dan *Sentiment Analysis* data keluhan masyarakat yang masuk. Manfaat *sentimen analysis* sangat penting untuk mengetahui sejauh mana data keluhan masyarakat terhadap pembangunan serta digunakan sebagai alat bantu untuk melihat respon masyarakat.

Mengingat jumlah data keluhan yang masuk begitu besar maka diperlukan sebuah proses analisa data yang mampu menangani hal ini. Salah satu alternatif yang tersedia saat ini adalah menggunakan analisa *big data*. Karakteristik data sumber dari analisa *big data* adalah data yang memiliki 3 karakteristik yaitu *volume*(ukuran data yang besar), *variety*(tipe datanya bervariasi dari data tidak terstruktur dan data terstruktur) dan *velocity*(transaksi data dalam jumlah yang besar).

1.2 Rumusan Permasalahan

Dalam penelitian ini inti permasalahan adalah belum adanya klasifikasi informasi dari masyarakat yang masuk pada media center terkait dengan proses pembangunan dan pelayanan yang dilaksanakan oleh Pemerintah Kota Surabaya.

1.3 Batasan Masalah

1. Data *media center* yang diteliti adalah *facebook* dan *twitter* sapawarga Pemerintah Kota Surabaya.
2. Proses klasifikasi hanya data teks berbahasa Indonesia.

1.4 Tujuan Dan Manfaat Penelitian

1.4.1 Tujuan

1. Penelitian ini memiliki tujuan untuk dapat mengklasifikasikan informasi yang masuk pada media center sebagai informasi positif atau negatif menggunakan metode *support vector machine* (SVM).
2. Memberikan masukan sebagai bahan acuan untuk filterisasi terhadap setiap keluhan pada *media center*.

1.4.2 Manfaat

1. Pengembangan aplikasi klasifikasi informasi subjektif terhadap data keluhan yang masuk.
2. Memberikan masukan sebagai bahan acuan untuk tim pengelola *media center* untuk mengidentifikasi data keluhan yang masuk.

1.5 Sistematika Penulisan

Dalam penulisan thesis ini, akan dibagi ke dalam beberapa bab, yaitu :

- BAB I PENDAHULUAN

Bab ini berisi latar belakang, tujuan, permasalahan, batasan permasalahan, sistematika pembahasan serta relevansi dan manfaat penelitian ini.

- BAB II KAJIAN PUSTAKA DAN LANDASAN TEORI

Berisi tentang kajian teoritis mengenai konsep dasar analisa *big data*, konsep *Mapreduce*, konsep *Algoritma Support Vector Machine* dan *sentiment analysis*. Disamping itu melakukan studi terhadap hasil-hasil penelitian sebelumnya serta literature pendukung lainnya.

- BAB III METODOLOGI

Membahas tentang perancangan sistem dan langkah-langkah dalam penelitian ini.

- BAB IV HASIL DAN PEMBAHASAN

Dalam bab ini dijelaskan hasil analisa hasil penelitian dan pembahasan.

- BAB V KESIMPULAN DAN SARAN

Berisikan kesimpulan-kesimpulan yang bisa diambil dari hasil penelitian ini serta saran-saran untuk penelitian selanjutnya.

BAB 2

KAJIAN PUSTAKA DAN LANDASAN TEORI

Pada bab ini dijelaskan dan dibahas teori-teori yang diperlukan guna menunjang serta menjadi acuan dalam penelitian ini. Dalam bab ini juga membahas penelitian sebelumnya sehingga menjadi acuan untuk menerapkan teori-teori yang sesuai yang diharapkan dapat mengarah ke tujuan yang ingin dicapai. Dasar-dasar teori yang akan dibahas yaitu *Sentiment Analysis*, *Media Center*, *big data*, *hadoop*, *MapReduce*, *classification data*, *Support Vector Machine*.

2.1 *Sentiment Analysis*

Sentiment analysis adalah studi komputasi mengenai sikap, emosi, pendapat, penilaian, padangan dari sekumpulan teks yang fokusnya adalah mengekstraksi, mengidentifikasi atau menemukan karakteristik sentimen dalam unit teks menggunakan metode *NLP* (*Natural Language Processing*), statistik atau *machine learning*.

Sentiment analysis merupakan proses klasifikasi dokumen tekstual ke dalam beberapa kelas seperti sentimen positif dan negatif serta besarnya pengaruh dan manfaat dari *sentiment analysis* menyebabkan penelitian ataupun aplikasi mengenai analisis sentimen. Saat ini perkembangan penelitian *sentiment analysis* mempunyai perkembangan yang sangat pesat bahkan di Amerika lebih dari 20 sampai 30 perusahaan memfokuskan pada layanan *sentiment analysis*. Pada dasarnya *sentiment analysis* merupakan klasifikasi, namun dalam implementasinya tidak mudah karena seperti proses klasifikasi biasa dikarenakan terkait penggunaan bahasa dimana terdapat ambiguitas dalam penggunaan kata, tidak adanya intonasi dalam sebuah teks, dan perkembangan dari bahasa itu sendiri.

Sentiment analysis bermanfaat juga dalam dunia usaha seperti melakukan analisa tentang sebuah produk yang dapat dilakukan secara cepat serta digunakan sebagai alat bantu untuk melihat respon konsumen terhadap produk tersebut, sehingga dapat membuat langkah-langkah strategis pada tahapan-tahapan berikutnya. Pada tugas akhir ini penelitian *sentiment analysis* terhadap keluhan masyarakat dengan menggunakan pendekatan dalam machine learning *Support Vector Machine* (SVM) dan dikhususkan pada dokumen teks bahasa Indonesia.

2.2 *Text Mining*

Text mining didefinisikan sebagai proses pengetahuan intensif di mana pengguna berinteraksi dengan koleksi dokumen dari waktu ke waktu dengan menggunakan seperangkat alat analisis (Romen Feldman, James Sanger 2007). Sama dengan data mining, *text mining* (pertambangan teks) dalam kerjanya yaitu mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi pola tertentu. *Text mining* sumber datanya adalah koleksi berbagai

dokumen. *Text mining* berawal dari berbagai penelitian tentang data mining sehingga merupakan hal yang biasa apabila ditemukan kesamaan arsitektur pengerjaan data mining maupun *text mining*. Misalnya *text mining* dan data mining mempunyai tahapan seperti *preprocessing*, *algoritma*, pola, hasil, serta tools atau alat bantu dan lain-lain. Dalam pengerjaannya *text mining* menggunakan banyak pola untuk mendapatkan hasil yang akurat sehingga untuk *text mining* pada pada tahapan *preprocessing* sangat penting karena tahapan ini untuk mengidentifikasi dan ekstraksi fitur representatif untuk dokumen bahasa, karena *preprocessing* bertanggung jawab untuk mengubah data dokumen. Tugas dari *text mining* adalah yaitu pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*).

Tujuan utama dari *text mining* adalah mendapatkan informasi yang berguna dari data yang diolah. Permasalahan yang dihadapi oleh *text mining* adalah data dalam jumlah yang besar, berdimensi tinggi, data yang berubah-ubah serta noise, tetapi yang membedakan *text mining* dan data mining adalah data dimana pada data mining menggunakan data terstruktur sedangkan pada *text mining* menggunakan data tidak terstruktur atau semistruktur sehingga merupakan tantangan dalam pengerjaan *text mining* dikarenakan struktur text yang tidak kompleks dan tidak lengkap, arti yang tidak jelas dan tidak standard, serta bahasa yang berbeda dan translasi yang tidak akurat. dalam *text mining* terdapat fitur-fitur pendukung yang sering digunakan antara lain :

1. **Character** Komponen individual yang ada dalam bagian *text mining* seperti huruf, angka, karakter spesial dan spasi, dan merupakan level paling tinggi dalam pembentukan *semantik feature* seperti kata, *term* dan konsep, tetapi pada umumnya *character-based* ini jarang digunakan pada teknik pemrosesan teks.
2. **Words** Diartikan kata-kata yang dipilih secara langsung dari dokumen asli yang menjadi dasar atau tingkatan dasar semantik, tetapi kadang fitur *word* terdapat didalam dokumen asli itu sendiri.
3. **Terms** Diartikan single word dan frasa *multi word* yang terpilih secara langsung dari korpus. Representasi term-based dari dokumen tersusun dari subset term dalam dokumen.
4. **Concept** Merupakan *feature* yang digenerate dari sebuah dokumen secara manual, *rule-based*, atau metodologi lain.

2.3 Riset Terkait

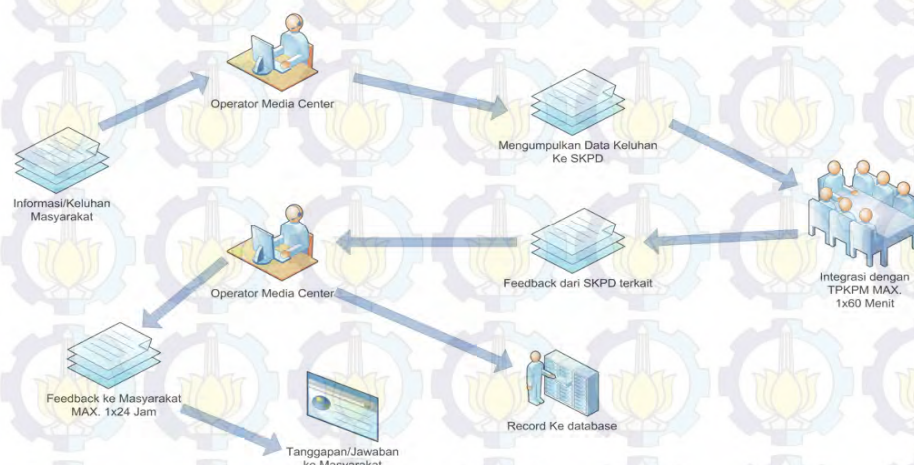
Penelitian yang dilakukan oleh (Edwin Lunando, 2007) yang mengangkat tentang media sosial Indonesia dengan *sarkasme Detection* menyimpulkan dalam media sosial Indonesia kecenderungan orang mengkritik sesuatu topik baik itu bernilai positif, negatif maupun netral masih menggunakan kata-kata sarkasme, untuk itu diperlukan fitur-fitur tambahan untuk mendeteksi *sentiment analysis* karena dari fitur-fitur tambahan seperti *unigram*, *sentiment score*, per-

nyataan dari kata-kata tersebut mampu memberikan informasi yang bernilai negatif dan lain-lain cukup efektif dalam mendeteksi *sarkasme*. Kekurangannya adalah penelitian ini hanya mendeteksi kata-kata *sarkasme* saja baik yang bernilai positif dan negatif dan topik yang dibahas sangat global sehingga hasil yang dicapai tidak menjadi sebuah tolak ukur.

Perbedaan yang dibuat penulis pada penelitian ini adalah penggunaan implementasi *sentiment analysis* yang dikhususkan pada topik pemerintahan sehingga dapat menghasilkan informasi yang menjadi tolak ukur dan menggunakan teks bahasa Indonesia saja. Metode yang digunakan adalah *support vector Machine(SVM)*.

2.4 Media Center

Media Center adalah sistem pelayanan informasi terintegrasi yang memberikan kesempatan bagi masyarakat Surabaya yang ingin berpartisipasi dalam perkembangan pembangunan kota Surabaya dan bentuk partisipasi masyarakat terwujud dalam keluhan, pengaduan, kritik, saran dan pertanyaan yang terkait dengan proses pembangunan dan pelayanan yang dilaksanakan oleh Pemerintah Kota Surabaya. Sebelum adanya *media center* keluhan masyarakat langsung ke Dinas Kominfo dimana menjadi pusat *media center*, tetapi dengan adanya media komunikasi media center maka masyarakat dapat memberikan keluhan, saran, kritikan melalui media komunikasi tanpa harus langsung ke Dinas Kominfo.



Gambar 2.1 SOP Media Center

Pada Gambar 1.2 dijelaskan bahwa sejak awal pembangunannya *Media Center* mempunyai tiga karakteristik yaitu responsif (merespon setiap data keluhan masyarakat yang masuk

kedalam Media Center) integratif (menggintegrasikan data keluhan masyarakat yang masuk ke Media Center) dan infomatif (memberikan informasi yang terupdate kepada masyarakat). Dalam sistem kerjanya media center menerima informasi atau keluhan masyarakat melalui media komunikasi kemudian operator mengumpulkan informasi tersebut dan memberikan kepada SKPD terkait setelah itu integrasi data dengan TPKPM maksimal 1x60 menit setelah itu feed back dari SKPD terkait ke operator *media center* dan *feed back* ke masyarakat maksimal 1x24 jam dan kemudian data disimpan ke dalam database.

2.5 Text Preprocessing

Struktur data yang baik dapat memudahkan proses komputerisasi secara otomatis. Pada *text mining*, informasi yang akan digali berisi informasi-informasi yang strukturnya sembarang (Noviah, 2013). Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk proses dalam data mining, yang biasanya akan menjadi nilai-nilai numerik. Proses ini sering disebut *Text Preprocessing*. Setelah data menjadi data terstruktur dan berupa nilai numerik maka data dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut. didalam preprocessing terdapat bagian-bagian dalam pengolahan teks seperti :

2.5.1 Pembersihan Dokumen(*Clansing*)

Proses membersihkan dokumen dari karakter-karakter yang tidak diperlukan untuk mengurangi noise seperti emotikon dan simbol-simbol.

2.5.2 Parsing

Proses untuk memecah teks bebas yang besar menjadi bagian-bagian yang disebut kalimat. Kalimat-kalimat yang dihasilkan kemudian dipecah lagi menjadi kata-kata melalui proses *tokenizing*.

2.5.3 Tokenizing

Proses memenggal setiap kata dalam teks, dan mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai 'z' yang diterima, sedangkan karakter selain huruf dihilangkan. Hasil dari proses tokenizing adalah kata-kata yang merupakan penyusun kalimat.

2.5.4 Stopword Removal

Proses penyaringan untuk menghilangkan kata yang tidak relevan pada hasil Tokenizing sebuah dokumen teks dengan cara membandingkannya dengan (*stopword list*) yang ada. Contoh dari *stopword* misalnya, kata sambung, artikel dan preposisi.

2.5.5 Stemming

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan confixes (kombinasi dari awalan dan akhiran) pada kata turunan. *Stemming* digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi Bahasa Indonesia yang baik dan benar. *Stemming* yang digunakan pada penelitian ini adalah *stemming* arifin-setiono, yang sudah banyak digunakan untuk proses *stemming* pada teks berbahasa Indonesia.

2.5.6 Normalisasi

Data yang sudah diakuisisi setelah melalui *preprocessing* atau normalisasi data tersebut perlu dilakukan penskalaan untuk sebelum dilakukan pelatihan terhadap data tersebut dinormalisasi dengan $mean=0$ dan $standard\ deviasi=1$ yang dijabarkan dengan rumus :

$$Nilai\ Baru = \frac{[Nilai\ Lama(Rata - Rata)]}{Standard\ Deviasi} \quad (2.1)$$

2.6 Big Data

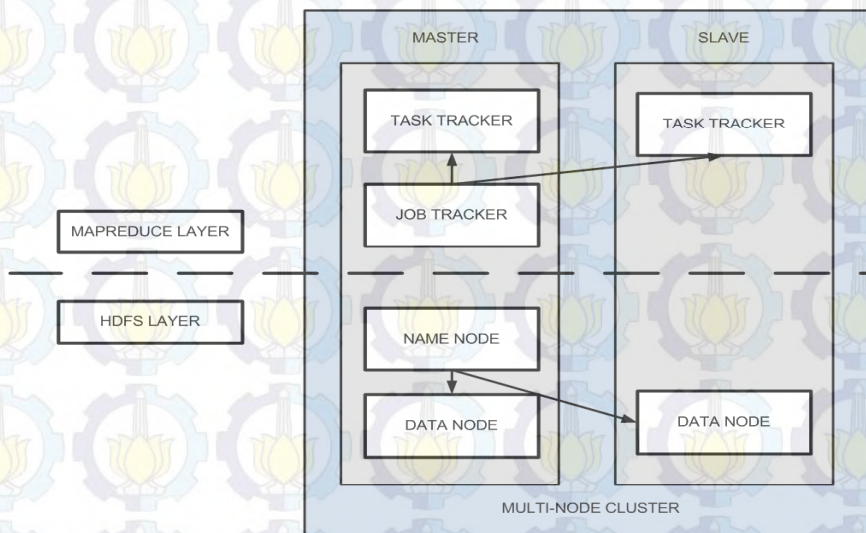
Saat ini proses pengolahan data baik dalam sistem pemerintahan maupun perusahaan swasta sudah menggunakan data center dan setiap bidang atau unit kerja sudah mempunyai data center dan hampir semuanya sudah terhubung antar satu dengan yang lainnya dan setiap hari datanya akan semakin bertambah dan semakin banyak variasi data yang yang disimpan serta jumlah transaksi data yang semakin besar maka diperlukan perangkat komputer yang sangat mahal dan membutuhkan tenaga IT yang sangat baik untuk mengoperasikannya (Judith Hurwith, 2013).

Untuk itu diperlukan proses analisa *Big Data* yang dalam pengertiannya sebagai pemecahan masalah ketika teknologi lama tidak lagi mampu melayani proses pengolahan data yang sangat besar. *Big data* mempunyai tiga karakteristik yaitu *volume* (ukuran data yang besar dan terdistribusi di banyak *server*), *variety* (tipe data bervariasi dari data terstruktur hingga dari tidak terstruktur), dan *velocity* (jumlah transaksi data yang besar sehingga perubahan ukuran data

juga akan semakin besar). Prinsip kerja big data yaitu tidak membuang atau menghapus sebuah data dikarenakan data tersebut menjadi penting dalam kurun waktu tertentu, proses data secara real-time dan mampu mengekstrasi dan transformasi sebuah data tanpa menghapus data awalnya.

2.7 Hadoop

Hadoop merupakan sebuah *software framework* teknologi terbaru berbasis *Java* dan sangat mudah didapatkan karena *hadoop* merupakan *software open source*. *Hadoop* diciptakan untuk pengolahan data yang sangat besar hingga *petabyte* dimana pengolahan data-data tersebut dilakukan dengan cara mendistribusikan data-data tersebut kedalam beberapa komputer yang telah di *cluster* dan komputer-komputer tersebut terhubung satu dengan lainnya.



Gambar 2.2 Hadoop distributed File System

Pada Gambar 2.2 dijelaskan Dalam perancangannya terdapat bagian seperti *Common Hadoop* yang fungsinya untuk menyediakan akses ke *filesystem* dan *Common Hadoop* berisi paket *file* dan skrip yang dibutuhkan *Hadoop* untuk memulai pekerjaannya. Paket ini menyediakan kode sumber, *document* dan bagian kontribusi yang cakupannya sangat besar dan waktu penjadwalan kerja yang efektif. *File system Hadoop* harus kompatibel karena wajib memberikan lokasi jaringan yang dipakai agar node dapat bekerja.

Salah satu contoh ketika cluster *Hadoop* kecil yang didalamnya terdapat sebuah *master node* dan beberapa *node* untuk bekerja atau lebih dikenal dengan *slave node*. *Master node* terdiri dari beberapa bagian yaitu *jobtracker*, *tasktracker*, *name node*, dan *data node*. *Node* untuk bekerja terdiri dari *data node* dan *tasktracker*, walaupun hanya untuk mendapatkan pekerja *node data*, dan hanya pekerja *node* menghitung.

Pada sistem *cluster* yang sangat besar, *file system HDFS* dikerjakan dengan *server name*

node diperuntukan pada host *indeks file system*, dan sebuah *name node* sekunder dapat menghasilkan snapshot dari struktur memori namenode, sehingga mencegah korupsi sistem file dan mengurangi hilangnya data. Demikian pula, *server jobtracker* dapat mengelola penjadwalan job secara mandiri. Dalam cluster *Hadoop MapReduce* mesin digunakan *mengcloud file system* alternatif, *name node* itu, *name node* sekunder dan arsitektur *data node* dari HDFS digantikan oleh setara *file* sistem-spesifik. Dalam sistem inti kerjanya *Hadoop* terdiri atas 2 bagian yaitu :

2.7.1 HDFS (*Hadoop Distributed File System*)

Merupakan sebuah *file system* yang fungsinya untuk menyimpan data yang sangat besar jumlahnya dengan cara mendistribusi data-data tersebut kedalam banyak komputer yang saling berhubungan satu dengan yang lainnya. Cara kerjanya yaitu *file* yang masuk kemudian dipecah-pecah dalam bentuk blok sebesar 64 MB atau bisa dikonfigurasi sendiri besarnya. Kemudian data direplikasi kedalam beberapa *node*(biasanya 3 *node*), dan disimpan dalam beberapa rak yang berbeda dengan tujuan agar menjaga reability dari HDFS.

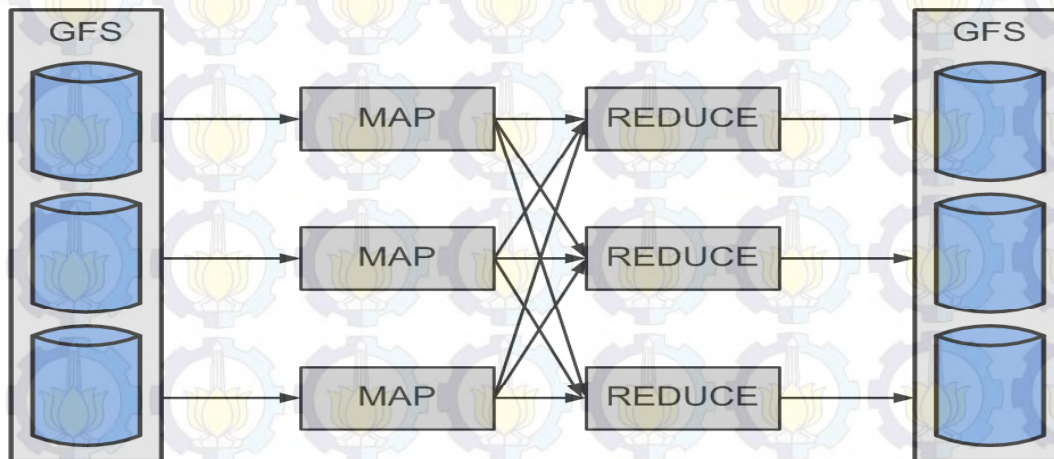
Untuk itu *file system* sangat membutuhkan *server* induk atau *master node* yang berfungsi untuk menyimpan metadata dari data yang ada di HDFS dan data-data tersebut disimpan dalam *server-server(datanode)* yang dapat diakses melalui *protokol* HTTP serta data *nodenya* saling terkait satu dengan lainnya untuk menjaga konsistensi datagan menggunakan *protokol* HTTP. Data *node* ini bisa saling berkomunikasi satu sama lain untuk menjaga konsistensi datadan memastikan proses replikasi data berjalan dengan baik.

Tetapi HDFS mempunyai kelemahan yaitu *master node* bersifat *Single Point of Failure* yang akan membuat data akan hilang apabila server *master node* mati. Walaupun dalam HDFS ada *secondary name node* tetapi tetapi *secondary name node* hanya menyimpan informasi terbaru dari struktur direktori pada *name node*. Untuk itu untuk mengatasi kelemahan yang ada maka dibuatkan *cloning* dari *server name node* ke beberapa *server* yang berbeda sehingga terjadi gangguan terhadap *name node* maka akan langsung digantikan oleh *cloningnya*.

Keuntungan dari HDFS adalah *jobtracker* dan *tasktracker* yang membuat jadwal dan peta serta mengurangi pekerjaan untuk *tasktrackers* pada lokasi-lokasi data. Sebagai contoh jika data pada *node* A (x, y, z) dan data yang terdapat *node* B (a, b, c). *jobtracker* akan jadwal *node* B untuk melakukan peta / mengurangi tugas pada (a, b, c) dan *node* A akan dijadwalkan untuk melakukan peta / mengurangi tugas pada (x, y, z). maka akan mengurangi jumlah lalu lintas yang berjalan di atas jaringan dan mencegah *transfer data* yang tidak perlu. *Hadoop* ketika digunakan dengan *file system* lain keunggulan ini tidak ada. Dan memberikan dampak yang signifikan terhadap waktu penyelesaian pekerjaan yang dapat ditunjukkan waktu data dijalankan dengan pekerjaan intensif.

2.7.2 MapReduce

Merupakan *software framework* yang digunakan untuk mendukung *distribusi computing* dengan menjalankan data yang sangat besar dan pertama kali diperkenalkan oleh *Google*.



Gambar 2.3 MapReduce

Pada Gambar 2.3 dijelaskan proses kerja MapReduce yang terdiri dari dua proses kerja yaitu :

1. **Map**

Sebuah proses ketika *master node* menerima masukan berupa data atau file, kemudian masukan tersebut dipecah menjadi beberapa bagian permasalahan yang kemudian didistribusikan ke *worker nodes*. *Worker nodes* ini akan memproses beberapa bagian permasalahan yang diterimanya untuk kemudian apabila *problem* tersebut sudah diselesaikan, maka akan dikembalikan ke *master node*.

2. **Reduce**

Sebuah proses ketika *Master node* menerima jawaban dari semua bagian permasalahan dari banyak data *nodes*, kemudian menggabungkan jawaban-jawaban tersebut menjadi satu jawaban besar untuk menghasilkan penyelesaian dari permasalahan utama. Keuntungan *MapReduce* adalah proses *map* dan *reduce* dijalankan secara terdistribusi. Setiap proses mapping sifatnya independen yang membuat proses dijalankan secara simultan dan paralel. Begitu juga dengan proses reducer dilakukan secara paralel pada waktu yang bersamaan, selama *output* dari operasi *mapping* mengirimkan *key value* yang sesuai dengan proses *reducernya*. Dalam proses *MapReduce* dapat diaplikasikan di *cluster server* dengan jumlah yang banyak sehingga dapat mengolah data dalam jumlah besar hanya dalam beberapa jam saja.

Dalam kerja *hadoop*, *mapreduce engine* ini terdiri dari satu *jobtracker* dan satu/banyak *tasktracker*. *JobTracker* merupakan server penerima *job* dari *client*, kemudian mendistri-

busikan *jobs* tersebut ke *tasktracker* yang akan mengerjakan *sub job* sesuai yang diperintahkan *jobtracker*. Sistem kerja ini mendekatkan pengolahan data dengan data itu sendiri, sehingga ini akan sangat signifikan mempercepat proses pengolahan data. Dalam kerjanya HDFS *file system* bukan hanya diperuntukan untuk *map/reduce* tetapi saat ini ada beberapa project lain yang related dengan *hadoop* yang dapat dijalankan diatas HDFS seperti *HBase, Pig, Hive*, dll.

2.8 Klasifikasi Data



Gambar 2.4 Model Klasifikasi

Pada Gambar 2.4 dijelaskan sistem klasifikasi data merupakan sebuah proses data-data yang masuk ke sebuah sistem pengolahan data baik itu data yang terstruktur maupun tidak terstruktur kemudian memetakan atau mengklasifikasikan setiap data kedalam salah satu dari beberapa kelas yang sudah didefinisikan.

Klasifikasi adalah sebuah proses untuk menemukan model yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui (*Tan et al*, 2004). Didalam klasifikasi diberikan sejumlah *record* yang dinamakan *training set*, yang terdiri dari beberapa atribut yang dapat berupa kontinyu ataupun kategori, salah satu atribut menunjukkan kelas untuk *record*.

2.8.1 Model Klasifikasi

1. Pemodelan Deskriptif Sebagai alat yang sifatnya untuk membedakan dan menjelaskan antara sebuah objek dengan kelas yang berbeda
2. Pemodelan Prediktif Sebagai alat yang digunakan untuk prediksi label kelas yang belum diketahui *recordnya*

2.8.2 Tujuan Klasifikasi

1. Menentukan sebuah model dari *training set* yang membedakan *record* ke dalam kategori atau kelas yang sesuai, dengan model tersebut kemudian digunakan untuk mengklasifikasikan *record* yang kelasnya belum diketahui sebelumnya pada testing set.

2. Mengambil keputusan dengan memprediksi suatu kasus, berdasarkan hasil klasifikasi yang diperoleh. Konsep pembuatan model dalam klasifikasi untuk mendapatkan model, kita harus melakukan analisis terhadap data latih (*training set*). Sedangkan data uji (*test set*) digunakan untuk mengetahui tingkat akurasi dari model yang telah dihasilkan. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai kelas dari suatu objek data.

2.8.3 Tahapan Proses klasifikasi data

1. Pembelajaran/Pembangunan Model

Tiap-tiap record pada data latih dianalisis berdasarkan nilai-nilai atributnya dengan menggunakan suatu algoritma klasifikasi untuk mendapatkan model.

2. Klasifikasi

Pada tahapan ini, data diuji digunakan untuk mengetahui tingkat akurasi dari model yang dihasilkan. Jika tingkat akurasi yang diperoleh sesuai dengan nilai yang ditentukan, maka model tersebut dapat digunakan untuk mengklasifikasikan *record-record* data baru yang belum pernah dilatihkan atau diuji sebelumnya.

Untuk meningkatkan akurasi dan efisiensi proses klasifikasi, terdapat beberapa langkah pemrosesan terhadap data, yaitu :

1. *Data Cleaning*

Data cleaning merupakan suatu pemrosesan terhadap data untuk menghilangkan *noise* dan penanganan terhadap *missing value* pada suatu *record*.

2. Analisis Relevansi

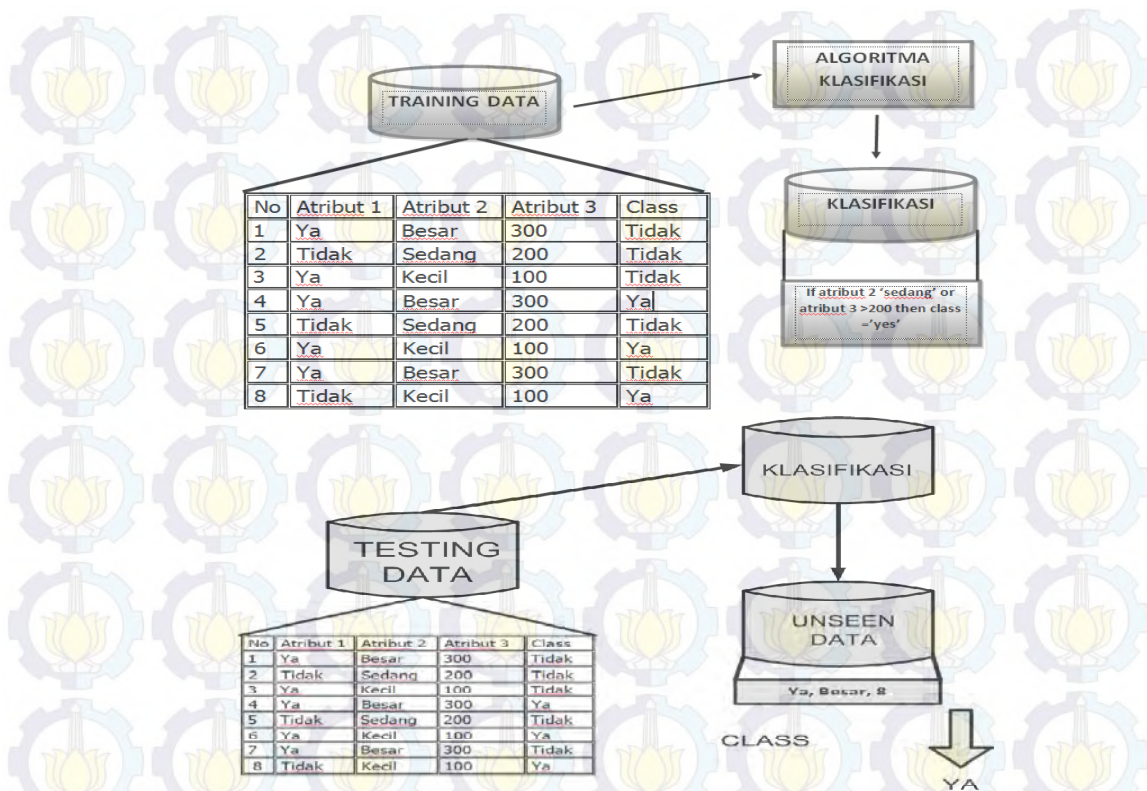
Pada tahap ini, dilakukan penghapusan terhadap atribut-atribut yang redundant ataupun kurang berkaitan dengan proses klasifikasi yang akan dilakukan. Analisis relevansi dapat meningkatkan efisiensi klasifikasi karena waktu yang diperlukan untuk pembelajaran lebih sedikit daripada proses pembelajaran terhadap data-data dengan atribut yang masih lengkap (masih terdapat redundansi).

3. Transformasi Data

Pada data dapat dilakukan generalisasi menjadi data level yang lebih tinggi. Misalnya dengan melakukan diskretisasi terhadap atribut dengan nilai kontinyu. Pembelajaran terhadap data hasil generalisasi dapat mengurangi kompleksitas pembelajaran yang harus dilakukan karena ukuran data yang harus diproses lebih kecil.

Pada Gambar 2.5 dijelaskan proses klasifikasi terdiri dari pembuatan model dan penggunaan model. Pembuatan model menguraikan sebuah set dari penentuan kelas-kelas sebagai :

1. Setiap tuple diasumsikan sudah mempunyai kelas yang dikenal seperti ditentukan oleh label kelas atribut.



Gambar 2.5 Proses Klasifikasi

2. Kumpulan tuple yang digunakan untuk membuat model disebut kumpulan pelatihan (*training set*)
3. Model direpresentasikan sebagai *classification rules*, *decision trees* atau formula matematika.

Penggunaan model menjelaskan pengklasifikasian masa yang akan datang atau objek yang belum diketahui, yaitu taksiran keakuratan dari model yang terdiri dari :

1. Label yang telah diketahui dari contoh tes dibandingkan dengan hasil klasifikasi dari model.
2. Nilai keakuratan adalah prosentase dari kumpulan contoh tes yang diklasifikasikan secara tepat oleh model.
3. Kumpulan tes tidak terikat pada kumpulan pelatihan.
4. Jika akurasi diterima, gunakan model untuk mengklasifikasikan data tuple yang label kelasnya belum diketahui.

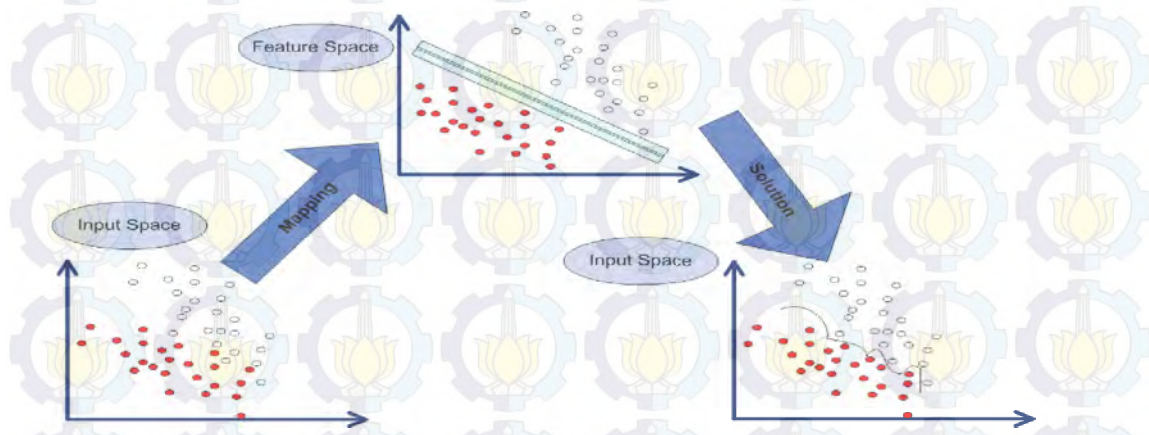
Untuk mengevaluasi performansi sebuah model yang dibangun oleh algoritma klasifikasi dapat dilakukan dengan menghitung jumlah dari test *record* yang diprediksi secara benar (akurasi) atau salah (*error rate*) oleh model tersebut. Akurasi dan *error rate* dapat dijelaskan dengan persamaan :

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Jumlah\ Total\ Prediksi} \quad (2.2)$$

$$ErrorRate = \frac{Jumlah\ Prediksi\ Salah}{Jumlah\ Total\ Prediksi} \quad (2.3)$$

2.9 Algoritma Support Vector Machine

Support Vector Machine (SVM) pertama kali dikembangkan oleh Boser, Guyon, dan Vapnik. Pada tahun 1992 ketika diadakan di *Annual Workshop on Computational Learning Theory*. *Support Vector Machine* (SVM) merupakan sistem pembelajaran terbimbing yang pengklasifikasiannya menggunakan ruang hipotesis berupa fungsi-fungsi linear dalam sebuah ruang fitur (*feature space*) berdimensi tinggi. (Budi Santosa, 2007)

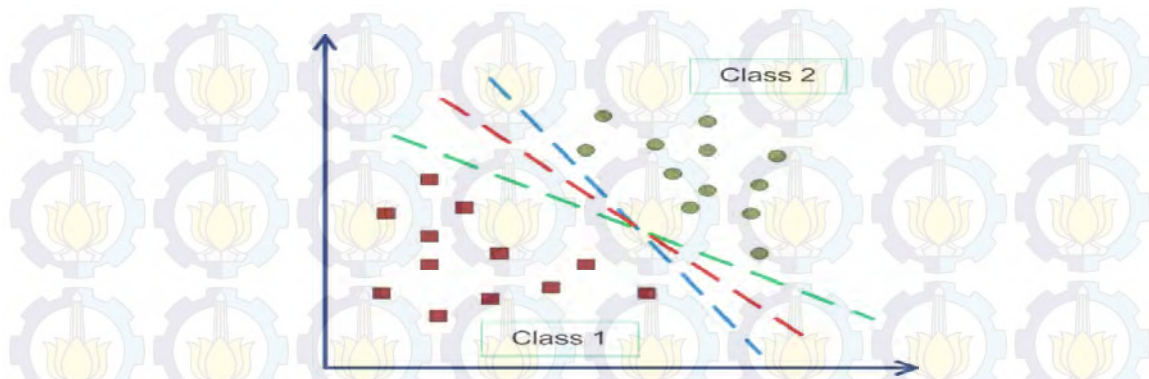


Gambar 2.6 Algoritma Support Vector Machine

Pada Gambar 2.6 dijelaskan metode dasar klasifikasi SVM dijelaskan dalam konsep SVM berusaha menemukan fungsi pemisah (*hyperplane*) terbaik diantara fungsi yang tidak terbatas jumlahnya. *Hyperplane* pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin *hyperplane* tersebut dan mencari titik maksimalnya. Pada awalnya prinsip kerja dari SVM yaitu mengklasifikasi secara linear (*linear classifier*), kemudian SVM dikembangkan sehingga dapat bekerja pada klasifikasi *non linear*. Formulasi optimasi SVM untuk masalah klasifikasi dibedakan menjadi dua kelas yaitu klasifikasi *linear* dan klasifikasi *non-linear*.

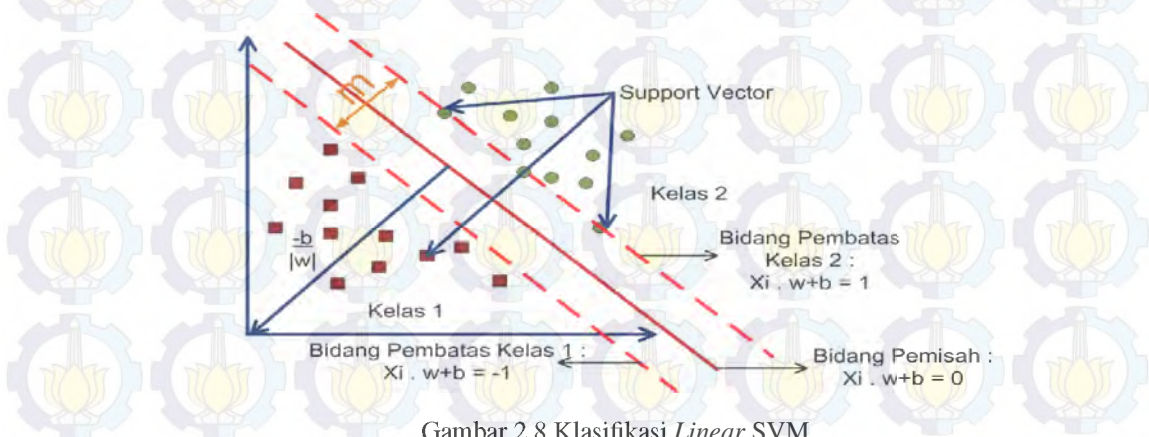
2.9.1 Klasifikasi Linear

Pada Gambar 2.7 dijelaskan dalam kerjanya SVM pada konsepnya secara sederhana diartikan sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah



Gambar 2.7 Klasifikasi *Linear SVM*

kelas pada *input space*. Dua kelas, +1 dan -1, beserta masing-masing pattern digambarkan dengan simbol kotak warna merah untuk *pattern* -1 dan simbol lingkaran warna hijau untuk *pattern* +1.



Gambar 2.8 Klasifikasi *Linear SVM*

Pada Gambar 2.8 dijelaskan dalam mengklasifikasi untuk mendapat hasil yang baik *hyperplane* digunakan untuk memisahkan menjadi dua kelas dengan mengukur margin *hyperplane* tersebut dan mencari titik maksimalnya, margin adalah jarak antara *hyperplane* terdekat dengan *pattern* terdekat dari masing-masing kelas dan *pattern* yang paling dekat dengan *hyperplane* disebut *support vector*. Seperti gambar dibawah ini garis tidak putus-putus yang terletak tepat di tengah-tengah kedua kelas. Sedangkan *support vector* tampak sebagai *pattern* yang berpotongan dengan garis putus-putus. Dari Gambar 2.7 bidang pemisah dapat dirumuskan :

m = jarak antara dua bidang

w = bidang normal

b = posisi relatif terhadap origin

jarak garis dirumuskan $wx+b=c$ dan ke origin adalah $(c-b)/|w|$

$$m = \frac{1 - b - (-1 - b)}{|w|} = \frac{2}{|w|} \quad (2.4)$$

Margin m dimaksimalkan dengan memenuhi konstrain 2 bidang pembatas yang sejajar dan data yang ada pada bidang pembatas disebut support vector. Bidang pembatas kelas pertama membatasi kelas pertama sedangkan bidang pembatas kelas kedua membatasi kelas kedua. sehingga diperoleh:

$$\begin{aligned} x_i \cdot w + b &\geq +1 \text{ for } y_i = +1 \\ x_i \cdot w + b &\leq -1 \text{ for } y_i = -1 \end{aligned} \quad (2.5)$$

Nilai maksimal margin harus memenuhi (2.4) dan (2.5) dan nilai b dan w dikalikan dengan sebuah konstanta yang akan menghasilkan nilai margin yang dikalikan dengan konstanta yang sama. Konstrain merupakan scaling constraint dengan dipenuhi rescaling b dan w . karena maksimalkan dan minimalkan w dirumuskan dalam pertidaksamaan (2.6).

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad (2.6)$$

Dengan mengalikan b dan sebuah konstanta, maka menghasilkan nilai m kemudian dikalikan dengan konstanta yang sama. Konstrain merupakan scaling constraint yang dipenuhi dengan rescaling b dan w . Maksimalkan $\frac{1}{|w|}$ = minimumkan $|w|^2$

Untuk mencari nilai margin terbesar untuk bidak pemisah terbaik dapat dirumuskan menjadi masalah optimasi konstrain, yaitu :

$$\begin{aligned} \min \frac{1}{2} |w|^2 \\ \text{s.t } y_i(x_i \cdot w + b) - 1 \geq 0 \end{aligned} \quad (2.7)$$

Dengan lebih mudah untuk menyelesaikan permasalahan optimasi konstrain dalam formulasinya dirubah kedalam formula lagrangian yang menggunakan lagrange multiplier yang diubah menjadi:

$$\min_{w,b}^{Lp}(w,b,a) \equiv \frac{1}{2} |w|^2 - \sum_{i=1}^n \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^n \alpha_i \quad (2.8)$$

Tambahan konstrain, $\alpha_i \geq 0$ (nilai dari koefisien lagrange). Dengan meminimumkan Lp terhadap w dan b , maka dari $\frac{\partial}{\partial b} Lp(w,b,a) = 0$ diperoleh (2.9) dan $\frac{\partial}{\partial w} Lp(w,b,a) = 0$ (2.10).

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad (2.9)$$

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad (2.10)$$

Sering kali vektor w bernilai besar bahkan tak terhingga. Tetapi nilai α_i , maka formula lagrangian Lp (primal problem) harus diubah kedalam dual problem LD dengan mendistribusikan

persamaan(2.10) ke L_D untuk memperoleh dual problem dengan kosntrain berbeda.

$$L_D(\alpha) \equiv \alpha - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j x_i x_j \quad (2.11)$$

maka permasalahan pencarian bidang pemisah terbaik dapat dirumuskan sebagai berikut :

$$\begin{aligned} \max_{\alpha} L_D &\equiv \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j x_i x_j \\ \text{s.t } \sum_{i=1}^n \alpha_i y_i &= 0, \alpha_i \geq 0 \end{aligned} \quad (2.12)$$

Untuk mendapatkan nilai α_i yang digunakan untuk menemukan w maka terdapat α_i untuk setiap data yang digunakan untuk pelatihan. data untuk pelatihan yang mempunyai nilai $\alpha_i > 0$ disebut support vector sedangkan yang memiliki nilai $\alpha_i = 0$ adalah sisanya. Maka fungsi untuk keputusan yang dihasilkan sangat dipengaruhi oleh *support vector*.

Formula pencarian bidang pemisah terbaik ini adalah permasalahan *quadratic programming*, sehingga nilai maksimum global dari α_i akan selalu dapat ditemukan setelah solusi permasalahan quadratic programming ditemukan (nilai α_i), maka kelas dari data pengujian x dapat ditentukan berdasarkan nilai dari fungsi keputusan:

$$f(X_d) = \sum_{i=1}^{ns} \alpha_i y_i x_i x_d + b \quad (2.13)$$

x_i adalah *support vector*,

ns = jumlah support vector

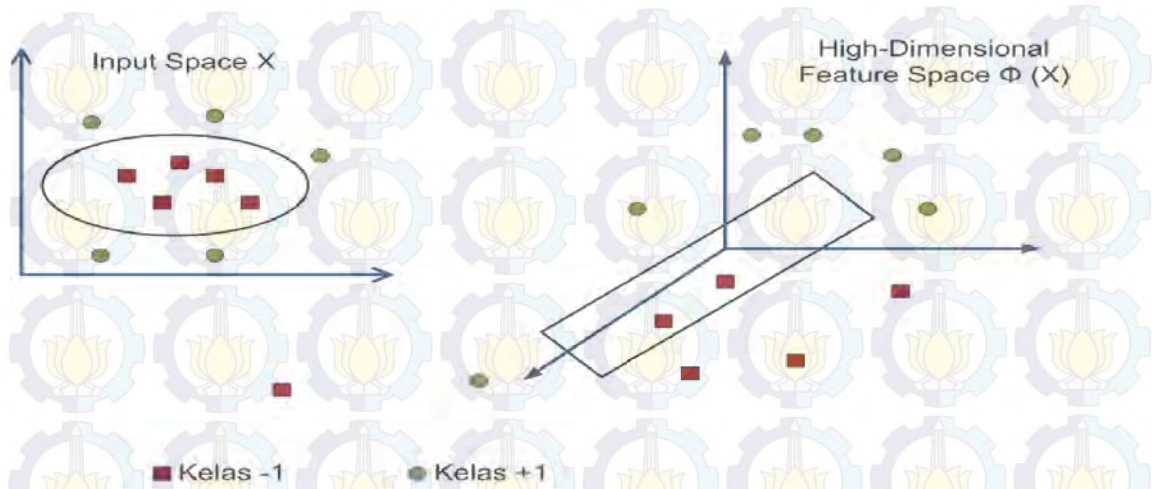
x_d adalah data yang akan diklasifikasikan.

2.9.2 Klasifikasi *Non Linear* SVM

Pada klasifikasi non-linear data yang berada dalam ruang sebuah *vector* awal harus dipindahkan ke ruang *vector* baru yang berdimensi lebih tinggi. Misal fungsi pemetaan dinotasikan sebagai ϕ . Pemetaan ini bertujuan untuk merepresentasikan data ke format yang linearly separable pada ruang *vector* baru.

Pada Gambar 2.9 dijelaskan bahwa proses optimisasi pada fase ini diperlukan perhitungan *dot product* dua buah contoh pada ruang *vector* baru. Dot product kedua buah *vector* x_i dan x_j dinotasikan sebagai $\phi x_i \cdot \phi x_j$. Nilai *dot product* kedua vector ini dapat dihitung secara tidak langsung, yaitu tanpa mengetahui proses transformasi ϕ .

Dalam melakukan klasifikasi yang tidak bisa dipisahkan secara linear(non linear)SVM harus memodifikasi formula untuk menyelesaikan permasalahan tersebut karena tidak ada solusi yang lain. Untuk itu kedua bidang pembatas(2.4)dan (2.5) harus dirubah sehingga tidak kaku



Gambar 2.9 Klasifikasi Non Linear SVM

untuk kondisi tertentu dengan adanya penambahan variabel ($\xi_i \geq 0 \forall_i : \xi = 0$) jika x_i diklasifikasi dengan benar) menjadi $x_i \cdot w + b \geq 1 - \xi_i$ untuk kelas 1 dan $x_i \cdot w + b \geq 1 + \xi_i$ untuk kelas 2. Untuk mencari bidang pemisah terbaik perlu penambahan ξ_i disebut soft margin hyperplane dan formula pencarian bidang pemisah terbaik berubah menjadi

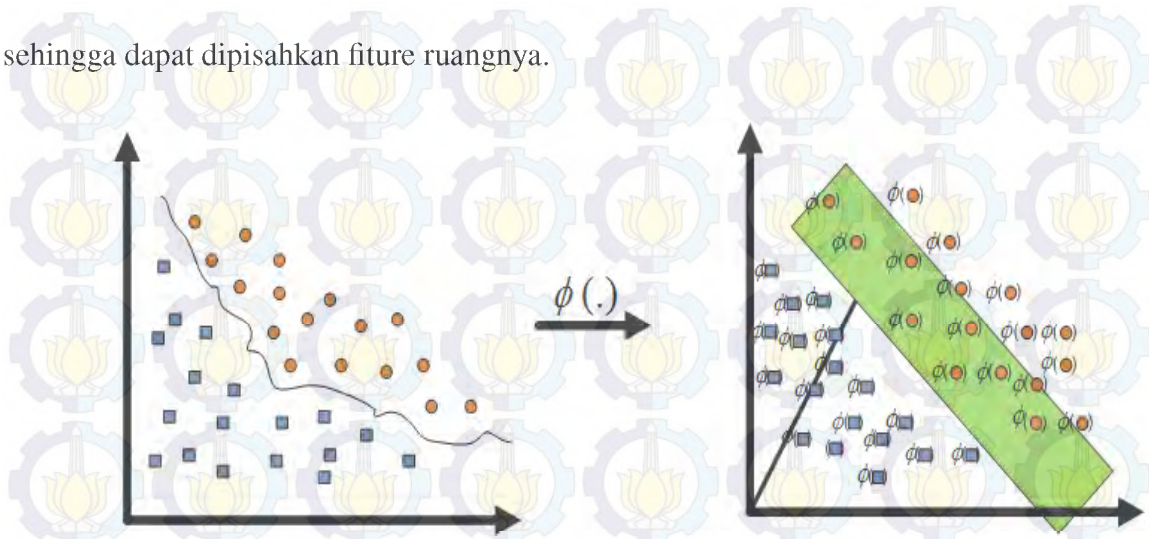
$$\begin{aligned} \min \frac{1}{2} \|w\|^2 + (\sum_{i=1}^n \xi_i) \\ \text{s.t. } y_i (w \cdot x_i + b) \geq -\xi_i \\ \xi \geq 0 \end{aligned} \quad (2.14)$$

C merupakan parameter yang menentukan besarnya penalti yang diakibatkan kesalahan klasifikasi data dan penentuan nilai biasanya dilakukan oleh pengguna. Permasalahan (2.14) memenuhi prinsip dari SRM, karena meminimumkan $\frac{1}{2} \|w\|^2$ ekuivalen dengan meminimumkan dimensi VC dan meminimumkan $C(\sum_{i=1}^n \xi_i)$ berarti meminimumkan error pada data pelatihan. Maka bentuk *primal problem* akan berubah menjadi :

$$\min_{w,b} L_p(w,b,a) \equiv \frac{1}{2} \|w\|^2 + C \left(\sum_{i=1}^n \xi_i \right) - \sum_{i=1}^n \alpha_i \{y_i(x_i \cdot w + b) - 1 + \xi_i\} - \sum_{i=1}^n \mu_i \xi_i \quad (2.15)$$

Diubahnya L_p ke dalam dual problem, maka akan menghasilkan sebuah persamaan yang sama dengan persamaan (2.10) maka pencarian bidang pemisah yang terbaik hampir sama dengan pemisahan secara linear. Pengubahan L_p ke dalam dual problem, menghasilkan formula yang sama dengan persamaan (2.6) sehingga pencarian bidang pemisah terbaik dilakukan menggunakan cara yang hampir sama dengan kasus dimana data dapat dipisahkan secara linier, tetapi rentang nilai α_i , tetapi rentang nilai $0 \leq \alpha_i \leq C$ serta instance yang memiliki nilai $\alpha_i = C$ disebut bounded support vector. Tetapi untuk klasifikasi data yang tidak dapat dipisahkan secara linear dapat dilakukan dengan metode transformasi data kedalam dimensi ruang fitur

sehingga dapat dipisahkan fitur ruangnya.



Gambar 2.10 Transformasi dari vektor input ke *feature space*

Pada Gambar 2.10 dijelaskan data dipetakan dengan menggunakan fungsi pemetaan atau transformasi caranya, data dipetakan dengan menggunakan fungsi pemetaan (transformasi) $x_k \rightarrow \phi(x_k)$ kedalam fitur ruang sehingga terdapat bidang pemisah yang mampu memisahkan data sesuai dengan kelasnya, contoh data set yang mempunyai dua atribut dan dua kelas yaitu positif dan negatif ke dalam *feature space* sehingga terdapat bidang pemisah yang dapat memisahkan data sesuai dengan kelasnya. Misalkan terdapat data set yang datanya memiliki dua atribut dan dua kelas yaitu kelas positif dan negatif. Data yang memiliki kelas positif adalah $(2,2), (2,-2), (-2,2), (-2,-2)$, dan data yang memiliki kelas negatif $(1,1), (1,-1), (-1,1), (-1,-1)$. Jika digambarkan dalam ruang dua dimensi maka data tidak bisa dipisahkan secara linear, oleh karena itu, digunakan fungsi transformasi sebagai berikut :

$$\phi(x_i, x_2) = \left\{ \begin{array}{l} \sqrt{x_1^2 + x_2^2} > 2 \rightarrow (4 - x_2 + |x_1 - x_2|, 4 - x_1 + |x_1 - x_2|) \\ \sqrt{x_1^2 + x_2^2} \leq 2 \rightarrow (x_1, x_2) \end{array} \right\} \quad (2.16)$$

Hasil dari transformasi data adalah Data sesudah transformasi $\{(2,2), (6,2), (6,6), (2,6)$ untuk kelas positif, dan $\{(1,1), (1,-1), (-1,1), (-1,-1)\}$ setelah itu dicari bidang pemisah terbaik dari data ini. Penggunaan fungsi transformasi $x_k \rightarrow \phi(x_k)$ maka nilai $w = \sum_{i=1}^{ns} \alpha_i y_i \phi(x_i)$ hasil pembelajaran adalah :

$$f(X_d) = \sum_{i=1}^{ns} \alpha_i y_i \phi(x_i) \phi(x_d) + b \quad (2.17)$$

Ruang fitur mampu memiliki jumlah fitur yang tidak terhingga dan sulit mengetahui fungsi transformasi yang tepat sehingga SVM perlu menggunakan "*kernel trick*" dari persamaan(2.15) dapat dilihat terdapat *dot product* $\phi(x_i) \phi(x_d)$. Apabila terdapat sebuah fungsi kernel

K Jika terdapat sebuah fungsi kernel K sehingga $K(x_i, x_d) = \phi(x_i) \cdot \phi(x_d)$ maka fungsi dari transformasi $\phi(x_k)$ diabaikan atau tidak perlu diketahui secara jelas. Fungsi yang dihasilkan dari pelatihan adalah :

$$f(x_d) = \sum_{i=1}^{ns} \alpha_i y_i K(x_i, x_d) + b \quad (x_i = \text{Support Vector}) \quad (2.18)$$

- Kernel Linear

$$K(x_i, x) = x_i^T x \quad (2.19)$$

- Polynomial Kernel

$$K(x_i, x) = (\gamma \cdot x_i^T x + r)^p, \gamma > 0 \quad (2.20)$$

- Radial Basis Function (RBF)

$$K(x_i, x) = \exp(-\gamma |x_i - x|^2), \gamma > 0 \quad (2.21)$$

- Sigmoid kernel

$$K(x_i, x) = \tanh(\gamma x_i^T x + r) \quad (2.22)$$

2.10 Precision, Recall, F-measure

Pada sistem informasi pencarian atau temu kembali dokumen, informasi akan memberikan sekumpulan dokumen sebagai jawaban dari query pengguna. dalam pencarian dokumen terdapat dua kategori yaitu relevant documents (dokumen yang relevan dengan query) dan retrieved documents (dokumen yang diterima pengguna). Kualitas data retrieval merupakan kualitas umum untuk mengukur kombinasi precision dan recall. Precision mengevaluasi kemampuan sistem temu kembali informasi untuk menemukan kembali data top-ranked yang paling relevan, dan didefinisikan sebagai persentase data yang dikembalikan yang benar-benar relevan terhadap query pengguna. Precision merupakan proporsi dari suatu set yang diperoleh yang relevan. Precision

$$\text{Precision} = \frac{\text{Relevant Document} \cap \text{Retrieved Document}}{\text{Retrieved Document}} \quad (2.23)$$

Dari persamaan 2.13 dijelaskan *Relevant* adalah jumlah dokumen yang *relevan*. Retrieved adalah jumlah dokumen yang dikembalikan atau diperoleh dari sistem kepada pengguna. *Recall* mengevaluasi kemampuan sistem temu kembali informasi untuk menemukan semua item yang relevan dari dalam koleksi data dan didefinisikan sebagai persentase data yang relevan terhadap *query* pengguna dan yang diterima. *Recall* merupakan proporsi dari semua hasil yang relevan di koleksi termasuk hasil yang diperoleh atau dikembalikan. Recall dapat dirumuskan

dalam persamaan (2.24) :

$$Recall = \frac{Relevant\ Documents \cap Retrieved\ Documents}{Relevant\ Documents} \quad (2.24)$$

Tabel 2.1 Tabel Kontingensi

Class	Actual Class		
		+	-
Predicted Class	+-	TP	FP
	-	FN	TN

$$Precision = \frac{TP}{(TP + FP)} \quad (2.25)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (2.26)$$

Dari Tabel 2.1 dijelaskan pada persamaan 2.25 dan 2.26 untuk mendapatkan nilai *precision* dan *recall*. Dengan TP adalah true positive yaitu jumlah dokumen yang di hasilkan aplikasi sama dengan jumlah dokumen yang dimasukkan sedangkan FP adalah false positive yaitu jumlah dokumen yang dianggap salah tetapi pada implementasinya dianggap benar(hasil yang tidak diinginkan) sedangkan FN adalah jumlah dokumen yang dianggap benar pada saat dimasukkan tetapi dianggap salah oleh pada implementasinya (missing value). Dari hasil perpaduan atau kombinasi antara *Recall* dan *Precision* disebut *harmonic mean*, biasa disebut *F-measure* yang mana dapat di formulasikan seperti persamaan seperti dibawah ini :

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} \quad (2.27)$$

F-measure pada prinsipnya digunakan pada sistem temu kembali informasi untuk mengukur klasifikasi pencarian dokumen serta performa query klasifikasi. sebelumnya *F-Measure* hanya difokuskan pada perhitungan nilai namun seiring perkembangannya *F-Measure* saat ini lebih menekankan pada kinerja *precision* dan *recall* sehingga bisa dilihat pada implementasinya secara keseluruhan.

BAB 3

METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

Pada Bab ini akan dijelaskan metode yang akan dilakukan pada penelitian ini, untuk menjawab permasalahan yang telah diajukan Bab 1 dalam perumusan masalah. Secara garis besar mirip dengan sistem *sentiment analysis* lainnya, sistem klasifikasi sentimen terdiri dari preprocessing, ekstraksi fitur dan klasifikasi. Pesan teks bahasa Indonesia dari media sosial seperti *Twitter* atau *Facebook*, orang cenderung menggunakan kata-kata tidak formal daripada yang formal seperti menggunakan angka untuk mengganti alfabet, karakter berulang vokal, dan menggunakan kata-kata informal yang umum untuk menggantikan kata-kata resmi. Untuk memproses kata-kata seperti itu, maka harus dilakukan tahapan-tahapan *preprocessing* seperti berikut:

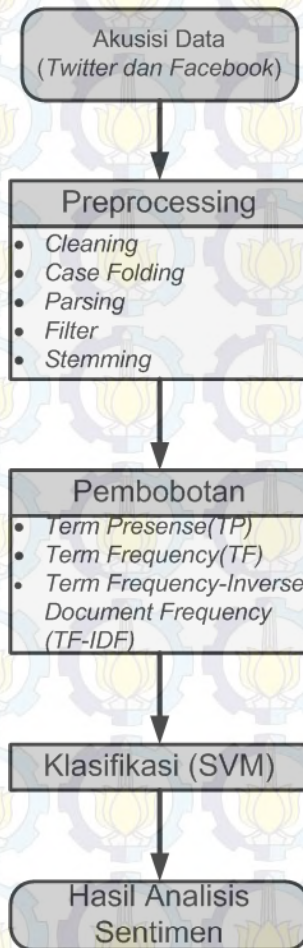
1. *Converse* karakter numerik ke dalam alfabet, seperti "du2k" menjadi "duduk" (duduk)
2. Hapus pengulangan vokal, seperti "aduuuh" menjadi "aduh"
3. Terjemahkan kata informal menjadi kata-kata resmi menggunakan kamus, seperti "cemu-ngudh" ke "Semangat" (semangat). Di sini, meskipun beberapa kata informal salah eja dari kata-kata formal, tetapi beberapa lainnya benar-benar berbeda leksikal dari kata-kata formal, karena itu strategi adalah untuk membangun sebuah kamus dan menggunakannya untuk menerjemahkan kata informal menjadi kata-kata formal.
4. Penggunaan kata-kata dalam bahasa Indonesia yang tidak baku dengan memakai kata-kata yang tidak sesuai dengan KBBI seperti "sudah" menjadi "sdh", "kemarin" menjadi "kmrn", "harapan" menjadi "hrpn".

Data teks berbahasa Indonesia diambil dari *twitter* (@SapawargaSby) dan *facebook* (Sapawarga Kota Surabaya). Data teks berupa komentar dari pengguna atau masyarakat kota Surabaya berupa data teks yang diakuisisi datanya, kemudian data teks tersebut diolah dan diklasifikasi menjadi opini positif, netral, atau negatif.

Dalam penelitian ini *case folding tokenisasi* dan *stemming* dilakukan dengan memecah-mecah kalimat menjadi kata dengan menghilangkan karakter-karakter selain dari alfabet dan menjadi term/kata dasar dan semua huruf kapital harus dirubah menjadi huruf kecil sehingga terjadi penyeragaman huruf dan token dapat diurutkan secara alfabet sehingga mempermudah dalam proses selanjutnya.

3.2 Perancangan Sistem

Pada Penelitian ini perancangan sistem menggambarkan kerangka atau pola dasar dari sistem yang akan dibangun. .



Gambar 3.1 Rancangan Sistem Klasifikasi

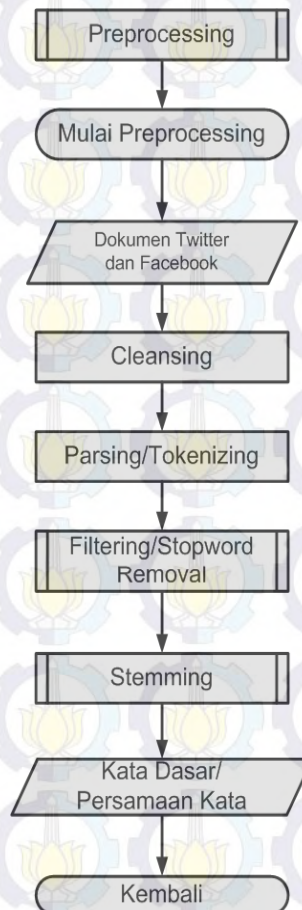
Pada Gambar 3.1 dijelaskan tahapan proses *sentiment analysis* pada sapawarga Kota Surabaya menunjukkan proses yang tahapan proses *sentimen analysis twitter* (@SapawargaSby) dan facebook(Sapawarga Kota Surabaya) melalui tahapan proses Akuisisi data kemudian tahapan berikutnya adalah preprocessing yaitu tahapan memproses data atau dokumen.

Pada tahapan awal ini data yang diakuisisi atau dikumpulkan dari situs jejaring sosial *Twitter* dan *Facebook* yang terhubung langsung melalui API (*Application Programing Interface*) dan menambahkan proses deteksi bahasa untuk mendapatkan data atau dokumen yang berbahasa indonesia.

Dalam tahapan preprocessing terdapat beberapa bagian yaitu *cleansing*, *case folding*, *parsing/tokenizing*, *filtering* kemudian tahapan *stemming* untuk mendapatkan kata dasar yang akan diklasifikasikan. tahapan berikutnya adalah pembobotan dengan menggunakan metode *term*

presense(TP) *term frequency*(TF) dan *term frequency-inverse documnet frequency*. Tahapan berikutnya proses klasifikasi menggunakan SVM.

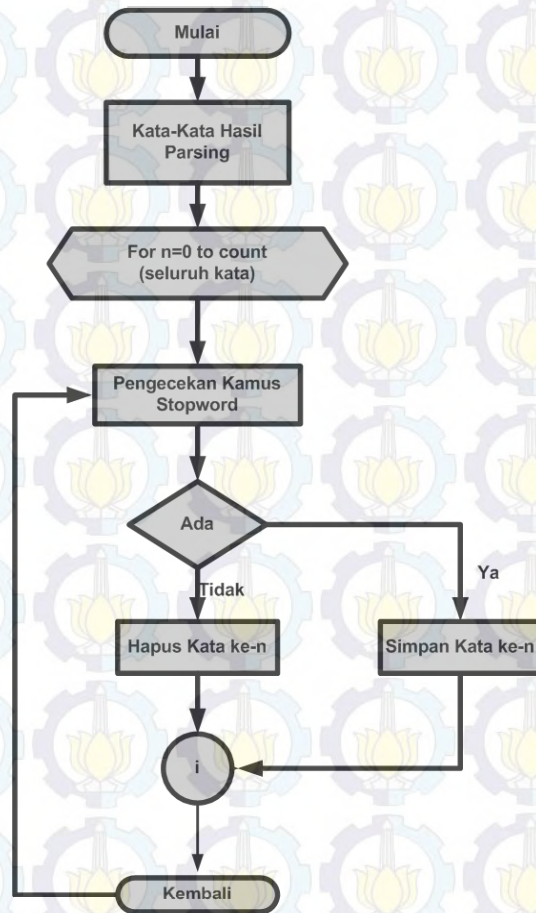
3.3 Diagram Alir Preprocessing



Gambar 3.2 Diagram Alur *Diagram Alir Preprocessing*

Pada Gambar 3.2 dijelaskan tahapan data atau dokumen yang diekstraksi dimasukkan kedalam sistem kemudian dilakukan proses pembersihan dokumen (*cleansing*) yang bertujuan untuk menghilangkan karakter-karakter kata-kata yang tidak diperlukan untuk mengurangi *noise* setelah itu dilakukan proses *case folding* yaitu menyeragamkan bentuk huruf-huruf mulai dari A sampai Z, huruf-huruf kapital diubah menjadi huruf kecil kemudian dilakukan proses *parsing tokenizing* yaitu membagi atau memecahkan dokumen-dokumen menjadi term-term berdasarkan *spasi stopword* kemudian dilakukan proses *filtering/stopword removal* untuk menyaring kata-kata atau dokumen-dokumen setelah itu dilakukan proses *stemming* untuk mendapatkan kata dasar atau persamaan kata. ini dilakukan berulang kali untuk mendapatkan kata dasar sesuai KBBI.

3.4 Diagram Alir *Filtering/Stopword Removal*



Gambar 3.3 Diagram Alir *Diagram Alir Preprocessing*

Pada Gambar 3.3 dijelaskan kata-kata yang sudah *diparsing/tokenizing* akan dihitung jumlah keseluruhan kata yang ada atau yang muncul dalam dokumen kemudian *filtering* atau pengecekan pada kamus untuk menghilangkan kata-kata yang tidak relevan yang ada pada dokumen sesuai kamus *stopword removal* jika tidak ditemukan kata-kata tersebut maka akan dihapus tetapi jika kata-kata tersebut ada dalam kamus maka akan disimpan ke dalam database.

3.5 Pembobotan

3.5.1 Term Weighting

Pada tahapan ini setiap dokumen dilakukan pembobotan untuk mendapatkan nilai data term/kata yang sudah melalui proses ekstraksi dan preprocessing. Dengan memakai metode TF-IDF untuk proses pembobotan, tahap pembobotan ini dokumen diubah menjadi sebuah vektor dengan banyak term yang dapat dikenali dari tahapan ekstraksi dokumen menggunakan metode TF-IDF

sebagai proses pembobotan. Pada tahapan ini kata dan simbol direpresentasi kedalam bentuk vektor dan TF-IDF, kata dan simbol direpresentasi ke dalam bentuk vektor dan tiap kata atau simbol dihitung sebagai satu fitur.

3.5.2 Term Presense(TP)

Pada tahapan ini fungsinya untuk menghitung jumlah kemunculan kata pada dokumen. Apabila jumlah sebuah term/kata probabilitas kemunculannya pada sebuah dokumen maka term/kata tersebut perlu mendapat query. Term frequency ini didasari pada aspek lokal pada TF-IDF monotonicity. Term frequency berfungsi untuk menentukan bobot dari ter/kata dalam sebuah dokumen yang berdasarkan frequency pada dokumen tersebut. Beberapa cara yang dilakukan untuk mendapatkan nilai TF antara lain :

- Raw TF TF sebuah term yang ada pada sebuah dokumen dihitung berdasarkan probabilitas kemunculan term/kata tersebut dalam dokumen.
- Logarithmic TF Nilai TF diperoleh dengan menggunakan fungsi logaritmik pada $tf = 1 + \log(tf)$. tf merupakan kemunculan term/kata dalam sebuah dokumen
- Binary TF TF di beri nilai boolean berdasarkan kemunculan term/kata pada sebuah dokumen, jika term tersebut tidak muncul dalam sebuah dokumen maka bernilai 0 sedangkan jika term tersebut muncul dalam sebuah dokumen maka term tersebut bernilai 1 sehingga dengan kemunculan term yang semakin banyak tidak akan berpengaruh.
- Augmented TF $tf = 0,5 + 0,5 \frac{tf}{\max(tf)}$ tf = jumlah kemunculan term dalam sebuah dokumen Dimana nilai tf adalah jumlah kemunculan term pada sebuah dokumen. $\max(tf)$ adalah jumlah kemunculan terbanyak term pada dokumen yang sama.

3.5.3 Inverse Document Frequency (IDF)

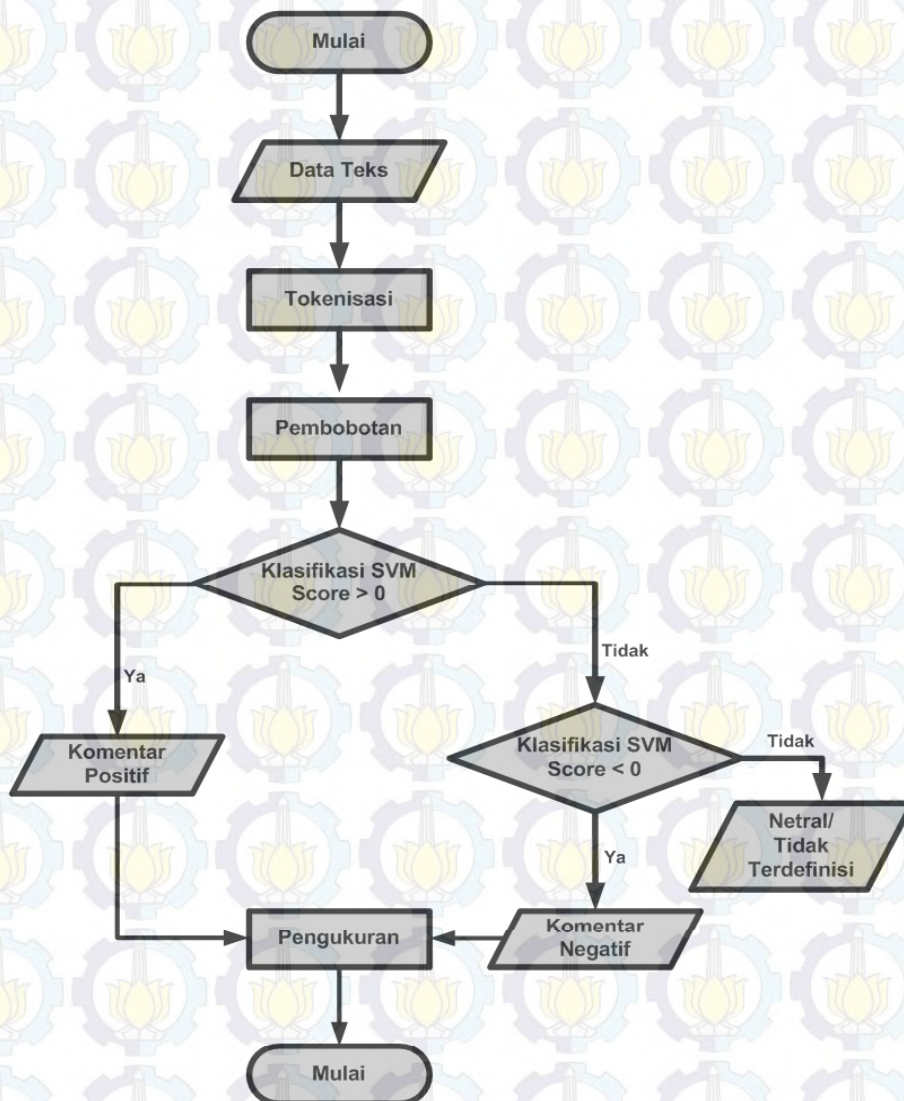
Merupakan jumlah dokumen yang berisikan term yang dicari dalam dokumen dataset. IDF sendiri biasa disebut Global weight yang fungsinya adalah mendefinisikan kontribusi dari term ke dalam dokumen. $idf = \log(\frac{N}{df})$ N : jumlah dokumen yang ada dalam kumpulan dokumen df : jumlah dokumen yang terdapat term. Hasil dari TF-IDF adalah dengan perkalian antara TD dan IDF.

3.5.4 TF-IDF

Metode pembobotan yang diintegrasikan dari term frequency (TF) dan inverse document frequency yang dijabarkan dengan rumus : $w(t, d) = tf(t, d) * idf$, metode ini fungsinya untuk mencari

representasi nilai dari kumpulan data yang ditraining yang hasilnya dibentuk vektor antara dokumen dan kata dan dicluster berdasarkan kesamaan antara dokumen dan term/kata.

3.6 Proses Klasifikasi(SVM)



Gambar 3.4 Diagram Alir Klasifikasi dengan SVM

Pada Gambar 3.4 dijelaskan data berupa kata-kata yang sudah melalui case folding, cleansing dan filtering akan ditokenizing kemudian dihitung pembobotannya, setelah pembobotan diklasifikasi apabila kata atau term tersebut bernilai < 0 maka kata tersebut dimasukkan dalam komentar negatif tetapi apabila kata tersebut $= 0$ maka kata tersebut dimasukkan kedalam komentar netral atau tidak terdefinisi dan apabila kata tersebut > 0 maka kata tersebut dikategori komentar positif.

Dalam klasifikasi sistem hanya melihat pada titik dan ruang daripada dokumen tersebut untuk tujuan pemodelan ruang vektor yang di gunakan kemudian memberikan setiap kata dalam dokumen yang akan diproses tersebut. serta bobot dari kata tersebut berdasarkan seberapa penting keberadaan kata tersebut dalam dokumen. SVM dalam proses klasifikasi ini berusaha untuk menemukan garis yang terbaik untuk membagi dua kelas kemudian diklasifikasikan dokumen uji berdasarkan pada sisi mana garis tersebut muncul.

Klasifikasi menggunakan machine learning SVM akan dimulai dengan mengubah teks menjadi data vektor kemudian vektor memiliki dua dimensi yaitu (*word id*) dan bobot. SVM dalam klasifikasi teks hanya pada titik dalam ruang daripada email atau dokumen sehingga model ruang akan memberikan setiap kata dalam sebuah dokumen id(dimensi dan sebuah bobot berdasarkan seberapa pentingnya dalam dokumen tersebut). Metode SVM dalam kerjanya mencoba untuk menemukan garis terbaik yang membagi dua kelas setelah itu melakukan klasifikasi dokumen yang diuji berdasarkan dari sisi mana dari garis tersebut yang muncul. SVM dalam klasifikasi menentukan garis terbaik yang memisahkan kedua kelas yang memiliki margin terbesar diantaranya.

Pada proses klasifikasi hasil pembobotan kata-kata akan langsung di hitung similaritas antara kata-kata atau dokumen-dokumen kemudian diurutkan hasil perhitungan dari similaritas dokumen-dokumen setelah itu hitung nilai-nilai pada setiap kategori dan menghitung probabilitas dokumen yang diuji pada masing-masing kategori dan mencari probabilitas yang paling besar setelah itu menentukan sentimen dokumen yang diuji sehingga menghasilkan data sentimen dokumen yang telah di uji. Penelitian menggunakan algoritma SVM. Untuk mendapatkan hasil klasifikasi terbaik, diujikan menggunakan tiga kelas sentimen kemudian membandingkan nilai dari tiga kelas tersebut.

$$\sum_i \alpha_i K(x_i, x) = \text{constant} \quad (3.1)$$

Dari persamaan (3.1) jika $K(x, y)$ lebih kecil dan nilai y bertambah dan lebih dari nilai x , pada setiap elemen dalam pengukuran penjumlahan dengan tingkat kedekatan titik uji x ke titik x_i dalam sebuah database yang sama atau sesuai. Penggunaan metode ini maka jumlah kernel dapat dipergunakan untuk pengukuran kedekatan relatif pada setiap titik uji dan titik data yang berasal dari satu atau kelompok himpunan lain. Pada penelitian kali ini diberikan yang dijabarkan dalam persamaan 3.2

$$D = \{(x_i, y_i) \mid X_i \in \{-1, 1\}\} \quad i^n = 1 \quad (3.2)$$

y_i adalah 1 atau -1, yang menunjukkan kelas mana dari titik x_i . x_i menyatakan vektor nyata p -dimensi. Untuk mencari hyperplane dari maksimum margin yang membagi nilai yang mempunyai nilai $y_i=1$ dari yang mempunyai $y_i=-1$. yang dijabarkan dengan persamaan 3.3

$$w \cdot x - b = 0 \quad (3.3)$$

Persamaan 3.3 menunjukkan *dot product* dan vektor w merupakan vektor normal yang tegak lurus dengan hyperplane tersebut, parameter $\frac{b}{\|w\|}$ menentukan offset hyperplane asal sepanjang vektor normal w . Pemilihan e dan b untuk maksimalkan jarak antara hyperplane secara paralel yang terbisah sejauh mungkin atau maksimalkan margin tetapi masih memisahkan data. dijabarkan pada persamaan 3.4

$$\begin{aligned} wx - b &= 1 \\ \text{dan} \\ wx - b &= -1 \end{aligned} \quad (3.4)$$

Apabila data yang dilatihkan terpisah secara linear maka bisa adakan pemilihan untuk dua hyperplane dari margin dengan cara menghilangkan atau tidak ada nilai antara hyperplane dan margin kemudian dimaksimalkan jaraknya dengan menggunakan persamaan $\frac{2}{\|w\|}$. cara ini untuk meminimalkan $\|w\|$. untuk menghindari titik data(i) jatuh kedalam margin maka perlu dirumuskan persamaan 3.5

$$\begin{aligned} wx_i - b &\geq 1 \quad \text{untuk } x_i \text{ sebagai kelas pertama} \\ \text{atau} \\ wx_i - b &\leq -1 \quad \text{untuk } x_i \text{ sebagai kelas kedua} \end{aligned} \quad (3.5)$$

yang dijelaskan dengan persamaan 3.6

$$y_i(w \cdot x_i - b) \geq 1, \text{ untuk semua } 1 \leq i \leq n \quad (3.6)$$

Sehingga untuk mendapatkan optimasi maka Minimalkan nilai w dan b ($\|w\|$) untuk setiap $i = 1, \dots, n$ yang dirumuskan

$$y_i(w \cdot x_i - b) \geq 1 \quad (3.7)$$

Apabila kelompok atau keluarga hyperplane ditemukan dengan membagi nilai maka

$$y_i(w \cdot x_i - b) - 1 \geq 0 \quad (3.8)$$

Untuk itu perlu dikirimkan semua nilai minimal dengan cara mengirimkan semua α_i untuk semua $+\infty$ sehingga nilai minimum ini akan dicapai oleh semua anggota kelompok atau keluarga dan tidak memilih yang terbaik untuk pemecahan masalah asal. Nilai constrain dinyatakan dengan persamaan 3.9 tujuan untuk mencari titik pegangan atau pelana

$$\max_{w,b} \max_a \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha [y_i(w \cdot x_i - b) - 1] \right\} \quad (3.9)$$

Dengan menggunakan persamaan 3.9 maka semua nilai dapat dipisahkan sebagai

$$y_i(w \cdot x_i - b) - 1 > 0 \quad (3.10)$$

Permasalahan dapat diselesaikan dengan teknik dan pemrograman kuadrat yang standar yang biasa disebut juga kombinasi linier dari vektor pelatihan yang dijabarkan

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad (3.11)$$

Sebagian α_i lebih besar dari nol dan x_i yang sesuai adalah vektor pendukung yang letaknya dimargin dan memenuhi $y_i(w \cdot x_i - b) = 1$. Vektor pendukung sendiri harus memenuhi

$$w \cdot x_i - b = 1/y_i = y_i \iff b = w \cdot x_i - y_i \quad (3.12)$$

Klasifikasi menggunakan bentuk unconstrained dual form bisa mendefinisikan bahwa hyperplane margin maksimum untuk itu klasifikasi berfungsi dari vektor pendukung dan data pelatihan yang berada pada margin. Klasifikasi dengan SVM bertujuan untuk mendapatkan hasil nilai dari dokumen yang diolah. Nilai $x_1 = 0, x_2 = 1$ dan kelas $y \in -1, 1$ maka persamaan untuk substitusi nilai x dan y diperoleh :

$$\begin{aligned} & \frac{1}{2} \alpha_1^2 - (\alpha_1 \alpha_2) \\ &= \frac{1}{2} \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \\ & \text{dengan konstrain :} \quad \alpha_1 - \alpha_2 = 0, \alpha_1 \geq 0, \alpha_2 \geq 0 \\ & \text{maka :} \quad \frac{1}{2} \alpha_1^2 - 2\alpha_1 = 0 \\ & \text{diperoleh} \quad \alpha_1 = \alpha_2 = 2 \end{aligned} \quad (3.13)$$

Untuk data yang lebih besar perlu dilakukan teknik matriks kernel untuk penyimpanan data pada memori, dengan data yang ukurannya besar maka dipakai metode 'working set' yang merupakan kumpulan variabel yang dioptimasi melalui current iteration. Fungsi kernel disini dijabarkan dengan :

$$k(x_i, x_j) = x_i \cdot x_j \quad (3.14)$$

Hyperplanes bias dan tidak bias pada perumusan klasifikasi dimaksudkan untuk bisa disederhanakan dengan alasan *hyperplane* melewati sistem koordinat yang asal biasa disebut *hyperplane unbiased* sedangkan *hyperplane* yang biasa atau pada umumnya tidak diharuskan melewati titik asal disebut bias. *Hyperplane* tidak bias dapat dilakukan dengan menetapkan nilai

$b = 0$ dalam permasalahan optimal primal yang dijabarkan

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad (3.15)$$

Metode decomposition salah satu metode pada SVM yang mengoptimasi permasalahan secara luas atau global dengan menggunakan data yang kecil atau sedikit pada saat dilakukan optimasi. Teknik decomposition biasa diimplementasikan dalam otasi matriks contohnya $\alpha = (\alpha_1, \dots, \alpha_i)^T$, $y = (y_1, \dots, y_i)^T$, $Q_{ij} = y_i y_j K(x_i, x_j)$ e adalah vektor yang jumlah elemennya sebanyak 1 (jumlah data pelatihan) dan semuanya bernilai 1 sehingga SVM dual problem dirumuskan :

$$\begin{aligned} \max e^T \alpha - \frac{1}{2} \alpha^T Q \alpha \\ \text{s.t. } 0 \leq \alpha_i \leq C, i = 1 \dots l \\ y^T \alpha = 0 \end{aligned} \quad (3.16)$$

Contoh vektor dibagi menjadi α_B dinyatakan sebagai variabel yang dimasukkan ke dalam working set sedangkan α_N merupakan variabel sisa. Matriks Q dapat dipartisi menjadi $Q = \begin{bmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{bmatrix}$ dan setiap bagiannya ditentukan oleh B dan N. Pada implementasinya SVM tidak terlalu sensitif terhadap imbalanced dataset karena hipotesa yang dihasilkan hanya dipengaruhi oleh sebagian data yang sudah menjadi support vector. Kelemahan dari SVM yaitu pada soft-margin hyperplane, misalnya jika parameter C kecil maka SVM cenderung melakukan klasifikasi data sebagai kelas mayoritas.

Untuk mengatasi permasalahan ini maka diperlukan bias untuk teknik klasifikasi sehingga lebih memperhatikan instance dari kelas minoritas dengan cara memberikan penalti lebih besar jika terjadi kesalahan pada proses klasifikasi kelas minoritas, dibandingkan jika salah dalam proses klasifikasi kelas mayoritas dan dirumuskan :

$$\begin{aligned} \min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C_+ \sum_{y_i = -1} \xi_i + C_- \sum_{y_i = 1} \xi_i \\ \text{s.t. } y_i (wx_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, \dots, n \end{aligned} \quad (3.17)$$

Teknik klasifikasi dapat dilakukan dengan pendekatan pada level data dengan melakukan undersampling pada kelas mayoritas dan oversampling pada kelas minoritas sehingga data yang dihasilkan menjadi balanced dataset. Pemilihan atribut sangat penting dalam klasifikasi menggunakan SVM karena atribut yang tidak penting pada dasarnya tidak mempengaruhi dalam proses klasifikasi untuk itu apabila atribut yang tidak penting dibuang maka akan meningkatkan efisiensi dari proses klasifikasi karena akan menjadi noise dalam proses klasifikasi.

Untuk mengetahui apakah atribut itu penting atau tidak maka dilakukan menggunakan

f-score yang merupakan teknik sederhana untuk mengukur tingkat diskriminasi dua buah vektor teknik sederhana untuk mengukur tingkat diskriminasi dua buah vektor bilangan desimal. Jika terdapat vektor data pelatihan $x_k, k = 1, \dots, m$. Jika jumlah data kelas positif adalah n_+ dan n_- adalah jumlah data kelas negatif. Maka F-score atribut ke i dirumuskan :

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} \left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} \left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (3.18)$$

dimana $\bar{x}_i, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ merupakan rata-rata nilai dari atribut keseluruhan data, data positif maupun negatif dan $x_{k,i}^{(+)}$ fitur ke i dari data kelas positif ke- k , $x_{k,i}^{(-)}$ fitur ke i dari data kelas negatif ke- k . Jika nilai $f(i)$ semakin besar maka atribut dapat dianggap lebih penting (diskriminatif).

Dari hasil klasifikasi hanya data yang berada pada perbatasan antar kelas yang mempengaruhi fungsi hasil keputusan pelatihan, untuk itu kandidat untuk support vector merupakan data memenuhi $y(x) f(x) \leq 1, y(x) f(x) > 1$ serta mendekati $y(x) f(x) = 1$. data yang dianggap kandidat support vector harus memenuhi :

$$y(x) f(x) \leq \beta + 1 \quad (3.19)$$

$$y(x) f(x) \leq 1 \quad (3.20)$$

Untuk penelitian kali ini menggunakan fungsi kernel polynomial maka model dari hyper-plane dijelaskan dengan persamaan :

$$f(x_d) = \sum_{i=1}^{N_{sv}} \alpha_i y_i (\gamma x_i^T x + r)^p + b \quad (3.21)$$

Keterangan :

N_{sv} := Jumlah Support Vector

α = Alpha, Pengali Lagrange

$i = 1, 2, 3, \dots, N_{sv}$

y = Label/Kelas dari data

b = bias

$(\gamma x_i^T x + r)^p + b$ = Persamaan polynomial kernel

Pada Tabel 3.1 dijelaskan proses klasifikasi pada data teks sapawarga Kota Surabaya dengan menggunakan *support vector machine* (SVM) yang dimasukkan ini dibagi menjadi dua data yaitu data pelatihan dan data pengujian yang dibagi dalam tujuh tahapan data set dengan variasi data pelatihan dan data pengujian yang berbeda-beda yang kemudian akan diproses.

Tabel 3.1 Tabel Data Set

Data	
Data Pelatihan	Data Pengujian
80%	20%
70%	30%
60%	40%
50%	50%
40%	60%
30%	70%
20%	80%

Secara umum data set tidak bisa dipisahkan secara linier, untuk itu data set harus di petakan ke feature space yang lebih tinggi melalui mapping sehingga didapatkan hyperplane pemisah yang lebih optimal dengan menggunakan fungsi kernel $\phi : x_i \rightarrow \phi(x)$ yang dijabarkan dalam persamaan $K(x_i, y_i) = \phi(x_i) \cdot \phi(x)$.

Langkah-langkah proses klasifikasi data dokumen teks pada *facebook* dan *twitter* sapa-warga Kota Surabaya sebagai berikut :

- Melakukan pembobotan dan normalisasi data pada masing-masing term/kata dengan menghitung frekuensi kemunculan masing-masing term/kata.
- Jumlah data pelatihan dan pengujian ditentukan beserta kelasnya menggunakan cross validation.
- Masukan matriks input (x) hasil dari pembobotan term/kata dan target input dari pelatihan tiga kelas.
- Pemetaan ke feature space yang lebih tinggi dengan tidak linear menggunakan fungsi kernel.
- Penentuan parameter terbaik $K(x_i, y_i) = \phi(x_i) \cdot \phi(x)^C = 2^{-5}, 2^{-3}, \dots, 2^{15}$ dan $K(x_i, y_i) = \phi(x_i) \cdot \phi(x)^C = 2^{-5}, 2^{-3}, \dots, 2^{15}$
- Model proses pelatihan berupa variabel-variabel hyperplane dari koordinat support vector(Nsv), weight(w) serta bias (b).
- Pengklasifikasian data pengujian pada model data pelatihan.
- Penentuan matriks data pengujian(x) yang semua fitur data yang digunakan pada proses pelatihan.

- Hasil pengujian berupa Confusion matrix Mendapatkan Confusion matrix hasil pengujian
- Perhitungan precision, recal dan F-Measure.

3.7 Precision, Recall dan F-Measure

Tabel 3.2 Tabel Confusion Matrix

Data Aktual	Data Prediksi		
	X	Y	Z
X	TP _x	FN _{xy}	FN _{xz}
Y	FN _{yx}	TP _y	FN _{yz}
Z	FN _{zx}	FN _{zy}	TP _z

Pada Tabel 3.2 dijelaskan tahapan klasifikasi yang diperoleh hasilnya kemudian hasil tersebut dilakukan proses perbandingan untuk mendapatkan nilai true positive, false positive, true negative dan false negative. Perhitungan dari multi kelas ini Hasil yang diperoleh pada tahap klasifikasi dilakukan proses perbandingan sehingga diperoleh empat nilai yaitu masing-masing adalah true positive, false positive, true negative dan false negative.

Untuk proses pengukuran akurasi multikelas, nilai True positive (TP_x) menjelaskan data yang menjadi bagian dari sebuah kelas teridentifikasi secara benar pada kelas tersebut. Nilai False positive (FN_{xy}, FN_{xz}, FN_{yx}, FN_{yz}, FN_{zx}, FN_{zy}) menjelaskan data yang bukan anggota suatu kelas teridentifikasi secara salah pada kelas tersebut. Nilai True negatif (TP_y, TP_z) menjelaskan data yang bukan anggota sebuah kelas tetapi teridentifikasi secara benar pada kelas tersebut (Y,Z). Nilai False negative (FN_{xy}, FN_{xz}, FN_{yx}, FN_{yz}, FN_{zx}, FN_{zy}) menjelaskan data yang bukan anggota sebuah kelas tetapi teridentifikasi secara salah pada kelas tersebut. gambaran perhitungan perbandingan dapat dilihat pada tabel

Nilai yang diperoleh dihitung precision, recall dan F-Measure untuk mengukur kecocokan terhadap data yang diuji sesuai dengan kelas yang ada persamaannya dirumuskan :

$$Precision = \frac{TP_x}{TP_x + FN_{xy} + FN_{xz}} \times 100\% \quad (3.22)$$

$$Recall = \frac{TP_x}{TP_x + FN_{yx} + FN_{zx}} \times 100\% \quad (3.23)$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} \quad (3.24)$$

BAB 4

HASIL DAN PEMBAHASAN

Bagian ini menjabarkan proses ekstraksi yang digunakan untuk mengklasifikasi kata-kata yang bernilai positif, negatif maupun netral dan studi kasus yang saya angkat adalah media sosial sapawarga Kota Surabaya *facebook* (Sapawarga Kota Surabaya) dan *Twitter* (@SapawargaSby). Akuisisi data teks berupa komentar keluhan masyarakat kemudian diekstraksi teks setelah itu diidentifikasi dan dikelompokkan menjadi kata-kata yang mengandung sentiment positif, sentimen negatif ataupun netral. *Support vector machine*(SVM) digunakan sebagai metode klasifikasi terhadap vector fitur teks yang diakuisisi lewat media sosial. Pada penelitian ini data yang diambil hanya data teks berupa komentar masyarakat melalui media sosial kemudian diklasifikasi berdasarkan kelas kata-kata tersebut.

4.1 Proses Akuisisi Data

Tabel 4.1 Akuisisi Data

No	Komentar
1	@ sapawarga @ suara surabaya tolong diinfokan cara pengurusan SITU atau SKTU di dinas apa ya untuk urusnya?"
2	@e100ss @SapawargaSby mhn di bantu pak...di daerah rmh saya listrik blm hidup mulai jam 12 mlm td....di daerah jl gresik,,trima kasih
3	@e100ss @SapawargaSby perbaikan lampu PJU d jl.mulyorejo,,terimakasih atas responnya... pic.twitter.com/x4WeftrRGo
4	@e100ss @SapawargaSby sy msk website dispendukcapil utk akta lahir online&dilink ke http://sapawarga.surabaya.go.id/lampid
5	@e100ss @SapawargaSby usul saja utk memberi rantai penghalang di pedestrian spy tdk dipakai spd motor nerobos saat macet
6	@roelswafa @SapawargaSby payah pegawai BPN surabaya1,waktu efektif kerja di pakai buat dangdutan,pelayanan kurang maksimal
7	@roelswafa: @SapawargaSby @e100ss lampu penerangan jalan raya sktr TL.jl.kenjaraan arah UNAIR C ada yang mati
8	@rolalisasi Sebagian pju jl ngagel jaya arah tl pertigaan ngagel jaya selatan mati. CC : @SapawargaSby
9	@SapawargaSby 4. simo tambahan. kemaren ada korban lagi, biasanya dlm 1 thn ada skitar 7-10 kecelakaan, klo gk jatuh, ya mati ditempat
10	@SapawargaSby ada info akun twitter utk polrestabes sby, pln, pdam, yg bersinggungan langsung dg warga. Penting lho supaya cepat terespon
11	@SapawargaSby apa ada E-KTP Corner di pusat perbelanjaan Surabaya? KTP saya sudah mati,ga ada waktu untuk ke Kecamatan
12	@SapawargaSby info free hotspot taman sulawesi ada gangguan pengunjung sore\ jadi agak sepi 0

Pada Tabel 4.1 dijelaskan dokumen yang diakuisisi yaitu dokumen berlabel teks media sosial sapawarga kota surabaya yaitu *facebook* dan *twitter* kemudian dokumen tersebut dimasukkan sebagai proses pembelajaran. Pada tahapan ini data masih utuh atau belum dibersihkan sehingga data teks yang akan diolah masih bercampur dengan karakter-karakter lainnya yang masih melekat pada data tersebut.

4.2 Preprocessing

Tahapan pengolahan data asli yang telah diakuisisi berupa data teks sebelum data tersebut diolah lebih lanjut. Tujuan dari preporcessing adalah menghilangkan noise, memperjelas fitur, mengkonversi data asli agar sesuai kebutuhan serta memperbesar atau memperkecil data. Dalam preprocessing sendiri terdapat beberapa tahapan antara lain :

4.2.1 Proses Pembersihan Dokumen(*Cleansing*)

Tabel 4.2 Pembersihan Dokumen

No	Komentar
1	suara surabaya tolong diinfokan cara pengurusan SITU atau SKTU di dinas apa ya untuk urusnya
2	mhn di bantu pak di daerah rmh saya listrik blm hidup mulai jam mlm td di daerah jl gresik trima kasih
3	perbaikan lampu PJU d jl mulyorejo terimakasih atas responnya
4	sy msk website dispendukcapil utk akta lahir onlinedilink ke
5	usul saja utk memberi rantai penghalang di pedestrian spy tdk dipakai spd motor nerobos saat macet
6	payah pegawai BPN surabaya1 waktu efektif kerja di pakai buat dangdutan,pelayanan kurang maksimal
7	lampu penerangan jalan raya sktr TL jl kenjaraan arah UNAIR C ada yang mati
8	rolalisasi Sebagian pju jl ngagel jaya arah tl pertigaan ngagel jaya selatan mati CC SapawargaSby
9	simo tambahan kemaren ada korban lagi biasanya dlm thn ada skitar kecelakaan klo gk jatuh ya mati ditempat
10	ada info akun twitter utk polrestabes sby pln pdam yg bersinggungan langsung dg warga Penting lho supaya cepat terespon
11	apa ada E KTP Corner di pusat perbelanjaan Surabaya KTP saya sudah mati ga ada waktu untuk ke Kecamatan
12	info free hotspot taman sulawesi ada gangguan pengunjung sore jadi agak sepi

Pada Tabel 4.2 dijelaskan dokumen yang sudah diakuisisi dibersihkan dari karakter-karakter seperti html, hastag, alamat situs(<http://www/situs.com>),username(@username),tanda baca seperti (.,?![\]/%;;<>()), angka(0,1,2,3,4,5,6,7,8,9,0) serta karakter-karakter lain selain alphabet. tujuan dari cleansing ini adalah mengurangi noise. Seperti pada tabel 4.2 dokumen teks yang sudah dibersihkan dari karakter-karakter yang tidak diperlukan. Tahapan ini sangat penting dilakukan agar kita mendapatkan data yang valid untuk diolah pada tahapan berikutnya.

4.2.2 Proses *Case Folding*

Pada Tabel 4.3 dijelaskan tahapan penyeragaman bentuk huruf dalam dokumen, huruf kapital diubah menjadi huruf kecil dan huruf-huruf diseragamkan dari A sampai Z dan selain huruf akan dihilangkan karena dianggap delimiter.

Tabel 4.3 Proses Case Folding

No	Komentar
1	suara surabaya tolong diinfokan cara pengurusan situ atau sktu di dinas apa ya untuk urusnya
2	mhn di bantu pak di daerah rmh saya listrik blm hidup mulai jam mlm td di daerah jl gresik trima kasih
3	perbaikan lampu pju d jl mulyorejo terimakasih atas responnya
4	sy msk website dispendukcapil utk akta lahir onlinedilink ke
5	usul saja utk memberi rantai penghalang di pedestrian spy tdk dipakai spd motor nerobos saat macet
6	payah pegawai bpn surabaya waktu efektif kerja di pakai buat dangdutan, pelayanan kurang maksimal
7	lampu penerangan jalan raya sktr tl jl kenjeraan arah unair c ada yang mati
8	rolalisasi sebagian pju jl ngagel jaya arah tl pertigaan ngagel jaya selatan mati cc sapawargasby
9	simo tambahan kemaren ada korban lagi biasanya dlm thn ada skitar kecelakaan klo gk jatuh ya mati ditempat
10	ada info akun twitter utk polrestabes sby pln pdam yg bersinggungan langsung dg warga penting lho supaya cepat terespon
11	apa ada e ktp corner di pusat perbelanjaan surabaya ktp saya sudah mati ga ada waktu untuk ke kecamatan
13	info free hotspot taman sulawesi ada gangguan pengunjung sore jadi agak sepi

4.2.3 Proses Parsing

Tabel 4.4 Parsing

D1	suara	surabaya	tolong	diinfokan	cara	pengurusan	situ	atau	sktu
	di	dinas	apa	ya	untuk	urusnya			
D2	mhn	di	bantu	pak	di	daerah	rmh	saya	listrik
	blm	hidup	mulai	jam	mlm	td	di	daerah	jl
	gresik	trima	kasih						
D3	usul	saja	utk	memberi	rantai	penghalang	di	pedestrian	spy
	tdk	dipakai	spd	motor	nerobos	saat	macet		
D4	payah	pegawai	bpn	surabaya	waktu	efektif	kerja	di	pakai
	buat	dangdutan	pelayanan	kurang	maksimal				
D5	sapawarga	lampu	penerangan	jalan	raya	sktr	tl	jl	kenjeraan
	arah	unair	c	ada	yang	mati			
D6	simo	tambahan	kemaren	ada	korban	lagi	biasanya	dlm	thn
	ada	skitar	kecelakaan	klo	gk	jatuh	ya	mati	ditempat

Pada Tabel 4.4 dijelaskan tahapan *parsing* atau disebut juga *tokenizing*. Pada tahapan sebuah dokumen kalimat dipecah-pecah menjadi kata-kata kemudian menganalisa terhadap kumpulan kata dengan memisahkan kata tersebut dan menentukan struktur sintaksis data tiap kata

tersebut.

D1 samapi D6 menunjukan dokumen yang di *parsing* atau *tokenizing* dan setiap kalimat dalam dokumen yang dipecah-pecah menjadi kata-kata sehingga menjadi bagian yang terpisah antara kata yang satu dengan kata yang lain.

4.2.4 Proses *Filtering/Stopword Removal*

Tabel 4.5 Filtering

D1	suara	surabaya	tolong			pengurusan		
			dinas			untuk	urusnya	
D2			bantu			daerah		
	listrik		hidup	mulai				
	daerah		gresik	trima	kasih			
D3	usul			memberi	rantai	penghalang		
			dipakai		motor	nerobos		macet
D4	payah	pegawai		surabaya	waktu	efektif	kerja	
	pakai			pelayanan	kurang	maksimal		
D5	sapawarga	lampu	penerangan	jalan	raya			
	kenjeraan						mati	
D6	tambahan	kemaren		korban		biasanya		
			kecelakaan			jatuh		mati

Pada Tabel 4.5 dijelaskan tahapan menghilangkan kata-kata yang tidak penting berdasarkan kamus stopwords seperti kata penghubung (yang, di, ke, ya) atau kata-kata yang tidak mempunyai makna. Untuk tahapan ini cukup menyulitkan dikarenakan komentar orang saat ini tidak memakai bahasa indonesia yang baku dan benar sehingga menyulitkan dalam melakukan penghapusan kata-kata, D1 sampai D6 menunjukkan dokumen kata-kata yang dihapus ditandai dengan kolom-kolom kata yang dikosongkan berikut. Proses Filtering juga menggunakan stopwords yaitu pencocokan kata-kata dengan kamus dalam hal ini memakai kamus bahasa Indonesia.

Pada Tabel 4.6 dijelaskan tahapan *filtering* setelah dibuang-kata-kata yang tidak penting dan tidak ada dalam *stopword* maka akan mendapatkan kata-kata yang akan diproses pada tahapan *tokenizing*.

Tabel 4.6 *Filtering 1*

D1	suara	surabaya	tolong	pengurusan	dinas	untuk	urusnya	
D2	bantu	daerah	listrik	hidup	mulai	daerah	gresik	trima
	kasih							
D3	usul	memberi	rantai	penghalang	dipakai	motor	nerobos	macet
D4	payah	pegawai	surabaya	waktu	efektif	kerja	pakai	pelayanan
	kurang	maksimal						
D5	sapawarga	lampu	penerangan	jalan	raya	kenjeraan	mati	
D6	tambahan	kemaren	korban	biasanya	kecelakaan	jatuh	mati	

4.2.5 Proses *Stemming*

Tabel 4.7 *Stemming*

D1	suara	surabaya	tolong	urus	dinas	untuk	urus		
D2	bantu	daerah	listrik	hidup	mulai	daerah	gresik	trima	kasih
D3	usul	beri	rantai	halang	pakai	motor	nerobos	saat	macet
D4	payah	pegawai	surabaya	waktu	efektif	kerja	pakai	layan	kurang
D5	sapawarga	lampu	terang	jalan	raya	kenjeraan	mati		
D6	simo	tambahan	kemaren	korban	biasa	celaka	jatuh	mati	tempat

Pada Tabel 4.7 dijelaskan tahapan mengubah kata-menjadi kata dasar dengan menghilangkan imbuhan (*affixes*) yaitu awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan yang ada didokumen akan dicocokkan dengan KBBI dan diubah sesuai pada aturan pada *stemming* (Arifin-Setiono) dan apabila dicek di kamus diubah sesuai aturan kata tidak ditemukan sebagai kata dasar maka kata tersebut dikembalikan kebentuknya semula dan dihitung sebagai kata dasar baru.

4.2.6 Proses Pembobotan

Pada Tabel 4.8 dijelaskan tahapan proses untuk menghitung dan informasi TF (*term frequency*), DF (*document frequency*) dan IDF (*inverse document frequency*) menghitung dokumen atau term ini berdasarkan frekuensi kemunculan term atau dokumen tersebut. Kata/term dihitung probabilitas kemunculan dalam satu dokumen (D1 sampai D6). untuk mendapatkan IDF digunakan persamaan dengan menggunakan persamaan)Setelah dilakukan proses preprocessing, selanjutnya dibuat tabel informasi dokumen yang berisi frekuensi term (TF), frekuensi dokumen (DF),

dan IDF dari masing-masing term. Kemudian dicari nilai $TF*IDF$ dari masing-masing term. Nilai IDF yang digunakan adalah nilai IDF yang didapatkan setelah proses pelatihan sistem. Nilai IDF dapat dirumuskan dalam persamaan sebagai berikut :

$$IDF = \log(D/DF) \quad (4.1)$$

contoh perhitungan pada baris pertama

$$= \log(6/1)$$

$$= 0.778$$

Tabel 4.8 Proses Pembobotan

No	Term/Kata	Dokumen						DF	IDF	TF*IDF					
		D1	D2	D3	D4	D5	D6			D1	D2	D3	D4	D5	D6
1	bantu	0	1	0	0	0	0	1	0.778	0.000	0.778	0.000	0.000	0.000	0.000
2	beri	0	0	1	0	0	0	1	0.778	0.000	0.000	0.778	0.000	0.000	0.000
3	celaka	0	0	0	0	0	1	1	0.778	0.000	0.000	0.000	0.000	0.000	0.778
4	daerah	0	2	0	0	0	0	2	0.477	0.000	0.477	0.000	0.000	0.000	0.000
5	dinas	1	0	0	0	0	0	1	0.778	0.778	0.000	0.000	0.000	0.000	0.000
6	efektif	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000
7	gresik	0	1	0	0	0	0	1	0.778	0.000	0.778	0.000	0.000	0.000	0.000
8	halang	0	0	1	0	0	0	1	0.778	0.000	0.000	0.778	0.000	0.000	0.000
9	hidup	0	1	0	0	0	0	1	0.778	0.000	0.778	0.000	0.000	0.000	0.000
10	jalan	0	0	0	0	1	0	1	0.778	0.000	0.000	0.000	0.000	0.778	0.000
11	jatuh	0	0	0	0	0	1	1	0.778	0.000	0.000	0.000	0.000	0.000	0.778
12	kasih	0	1	0	0	0	0	1	0.778	0.000	0.778	0.000	0.000	0.000	0.000
13	kemaren	0	0	0	0	0	1	1	0.778	0.000	0.000	0.000	0.000	0.000	0.778
14	kenjeraan	0	0	0	0	1	0	1	0.778	0.000	0.000	0.000	0.000	0.778	0.000
15	kerja	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000
16	korban	0	0	0	0	0	1	1	0.778	0.000	0.000	0.000	0.000	0.000	0.778
17	kurang	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000
18	lampu	0	0	0	0	1	0	1	0.778	0.000	0.000	0.000	0.000	0.778	0.000
19	layan	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000
20	listrik	0	1	0	0	0	0	1	0.778	0.000	0.778	0.000	0.000	0.000	0.000
21	macet	0	0	1	0	0	0	1	0.778	0.000	0.000	0.778	0.000	0.000	0.000
22	maksimal	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000

Lanjutan

23	mati	0	0	0	0	1	1	2	0.477	0.000	0.000	0.000	0.000	0.477	0.477
24	motor	0	0	1	0	0	0	1	0.778	0.000	0.000	0.778	0.000	0.000	0.000
25	mulai	0	1	0	0	0	0	1	0.778	0.000	0.778	0.000	0.000	0.000	0.000
26	nerobos	0	0	1	0	0	0	1	0.778	0.000	0.000	0.778	0.000	0.000	0.000
27	pakai	0	0	1	1	0	0	1	0.477	0.000	0.000	0.477	0.477	0.000	0.000
28	payah	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000
29	pegawai	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000
30	rantai	0	0	1	0	0	0	1	0.778	0.000	0.000	0.778	0.000	0.000	0.000
31	sapawarga	0	0	0	0	1	0	1	0.778	0.000	0.000	0.000	0.000	0.778	0.000
32	simo	0	0	0	0	0	1	1	0.778	0.000	0.000	0.000	0.000	0.000	0.778
33	suara	1	0	0	0	0	0	1	0.778	0.778	0.000	0.000	0.000	0.000	0.000
34	surabaya	1	0	0	1	0	0	2	0.477	0.477	0.000	0.477	0.000	0.000	0.000
35	tambahan	0	0	0	0	0	1	1	0.778	0.000	0.000	0.000	0.000	0.000	0.778
36	tempat	0	0	0	0	0	1	1	0.778	0.000	0.000	0.000	0.000	0.000	0.778
37	terang	0	0	0	0	1	0	1	0.778	0.000	0.000	0.000	0.000	0.778	0.000
38	tolong	1	0	0	0	0	0	1	0.778	0.778	0.000	0.000	0.000	0.000	0.000
39	trima	0	1	0	0	0	0	1	0.778	0.000	0.778	0.000	0.000	0.000	0.000
40	urus	2	0	0	0	0	0	2	0.477	0.477	0.000	0.000	0.000	0.000	0.000
41	usul	0	0	1	0	0	0	1	0.778	0.000	0.000	0.778	0.000	0.000	0.000
42	waktu	0	0	0	1	0	0	1	0.778	0.000	0.000	0.000	0.778	0.000	0.000

4.3 Klasifikasi dengan *Support Vector machine (SVM)*

Tabel 4.9 Split Data Pelatihan dan Data Pengujian

Perbandingan Data		Jumlah Data	
Data Pelatihan	Data Pengujian	Data Pelatihan	Data Pengujian
80%	20%	744	186
70%	30%	651	379
60%	40%	558	372
50%	50%	465	465
40%	60%	372	558
30%	70%	279	651
20%	80%	186	744

Pada Tabel 4.9 dijelaskan dari hasil akuisisi data kemudian lewat *preprocessing* data yang ada sebanyak 930 yang dibagi dalam 3 kelas sentiment yaitu positif(1) netral(0) dan negatif(-1). Data yang sudah dinormalisasi sebelum dimasukkan ke mesin klasifikasi, data tersebut dibagi menjadi 2 yaitu data pelatihan dan data pengujian dengan menggunakan metode *cross validation*. pada pengujian ini dilakukan sebanyak 7 kali dengan inputkan data training dan data testing yang berbeda.

Data yang akan di dilatihkan maupun di diujikan telah dibagi dan ditiap percobaannya menggunakan mesin pendukung vektor (SVM) Data vektor fitur yang telah dibagi ditiap percobaannya dilakukan klasifikasi menggunakan mesin pendukung vektor (SVM) dengan fungsi *polynomial kernel* yang memetakan data yang *non-linear* sehingga mendapat data set baru model pembelajaran pada tiap-tiap percobaan.

Paling utama dalam melakukan percobaan adalah pemilihan parameter dari mesin pembelajaran SVM dan fungsi *polynomial kernel* hal yang penting perlu dilakukan adalah pemilihan parameter dari mesin klasifikasi (SVM) dan fungsi *polykerner* yaitu parameter C dan γ (gamma) dengan *k-fold cross validation*. Parameter C dan γ dapat ditentukan yaitu dengan memberikan nilai parameter C yang pengaturannya pada interval 2^{-5} sampai dengan 2^{15} dan untuk parameter γ (gamma) pada interval 2^{-15} sampai dengan 2^3 . Teknik *cross-validation* digunakan untuk menentukan nilai terbaik parameter C pada interval nilai yang telah ditentukan untuk kedua parameter tersebut.

Hasil dari model pembelajaran diklasifikasi dengan percobaan pengujian sebanyak 7 kali dimana setiap satu kali pengujian diperoleh matrix dengan ukuran 3x3 sebagai representatif kelas aktual dan kelas prediksi. Setelah diperoleh hasil model dari pembelajaran mesin klasifikasi *support vector machine*(SVM) dilakukan percobaan pengujian menggunakan data uji yang telah di split sebanyak 7 kali tabel 4.5 dimana untuk hasil pengujian klasifikasi diperoleh suatu matrix dengan ukuran 3x3 sebagai representasikan sebuah kelas aktual dan kelas prediksi. Hasil dari pada model pembelajaran tersebut dilakukan pengujian dengan menggunakan data baru yang belum dilakukan pembelajaran sebelumnya.

Tabel 4.10 *Confusion Matrix*

Data Aktual	Data Prediksi		
	Positif	Netral	Negatif
Positif	TP	PN	FP
Netral	NP	TN	NNg
Negatif	FN	NgN	TNg

Pada Tabel 4.10 dijelaskan *Confusion matrix* merupakan hasil prediksi menggunakan mesin klasifikasi SVM yang diukur performa dari tiap-tiap kelas dengan cara menghitung *preci-*

sion, recall dan *F-Measure*. Keterangan :

TP = Kelas kata terprediksi benar bernilai positif

PN = Kelas kata positif terprediksi netral

FP = Kelas kata positif terprediksi negatif

NP = Kelas kata netral terprediksi sebagai kelas kata positif

TN = Kelas kata terprediksi netral

NNg = Kelas kata netral terprediksi negatif

FN = Kelas kata negatif terprediksi kata positif

NNg = Kelas Kata negatif terprediksi netral

TNg = Kelas Kata Negatif terprediksi negatif

Precision digunakan untuk menghitung akurasi kelas yang terprediksi sesuai dengan kelas aktual untuk hasil akurasi. Untuk pengukuran *Precision* digunakan persamaan berikut :

$$Precision = \frac{True\ Positif}{(True\ Positif + False\ Positif)} \quad (4.2)$$

Perhitungan *precision* tiap-tiap kelas kata digunakan persamaan :

$$Positif = \frac{TP}{(TP + NP + FN)} \quad (4.3)$$

$$Netral = \frac{TN}{(TN + PN + NgN)} \quad (4.4)$$

$$Negatif = \frac{TNg}{(TNg + FP + NNg)} \quad (4.5)$$

Recall digunakan untuk mengukur sensitifitas pengukuran terhadap dataset atau kemampuan prediksi sistem sesuai dengan tingkat kebenaran untuk memanggil dokumen yang relevan. untuk pengukuran recall digunakan persamaan :

$$Recall = \frac{True\ Positif}{(True\ Positif + False\ Negatif)} \quad (4.6)$$

Perhitungan *recall* tiap-tiap kelas kata digunakan persamaan :

$$Positif = \frac{TP}{(TP + PN + FP)} \quad (4.7)$$

$$Netral = \frac{TN}{(TN + NP + NNg)} \quad (4.8)$$

$$Negatif = \frac{TNg}{(TNg + FN + TNg)} \quad (4.9)$$

4.3.1 Percobaan Dengan Data Pelatihan 80% dan Data Pengujian 20%

Tabel 4.11 Hasil Percobaan Pertama

Data Aktual	Data Prediksi			Precision	Recall	F-Measure	Kelas
	Positif	Netral	Negatif				
Positif	8	92	0	0.444	0.080	0.136	Positif
Netral	10	596	0	0.821	0.983	0.895	Netral
Negatif	0	38	0	0.000	0.000	0.000	Negatif

Pada Tabel 4.11 dijelaskan pada percobaan pertama digunakan set data yang telah displit secara *random* diperoleh data pelatihan 80% dan data pengujian 20% yang akan diklasifikasikan menggunakan model pembelajaran SVM maka hasil yang didapatkan berupa *confusion matrix*.

Perhitungan *precision* hasil berdasarkan Tabel 4.11 dihitung menggunakan persamaan 4.2, 4.3, dan 4.4 dan *recall* menggunakan persamaan 4.6, 4.7 dan 4.8 serta *F-Measure* berdasarkan persamaan 4.9 sebagai berikut :

Kelas positif

$$Precision = \frac{TP}{(TP+NP+FN)} = \frac{8}{(8+10+0)} = 0.444$$

$$Recall = \frac{TP}{(TP+PN+FP)} = \frac{8}{(8+92+0)} = 0.080$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.444 * 0.080}{(0.444) + (0.080)} = 0.136$$

Kelas netral

$$Precision = \frac{TN}{(TN+PN+NgN)} = \frac{596}{(596+92+38)} = 0.821$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{596}{(596+10+0)} = 0.983$$

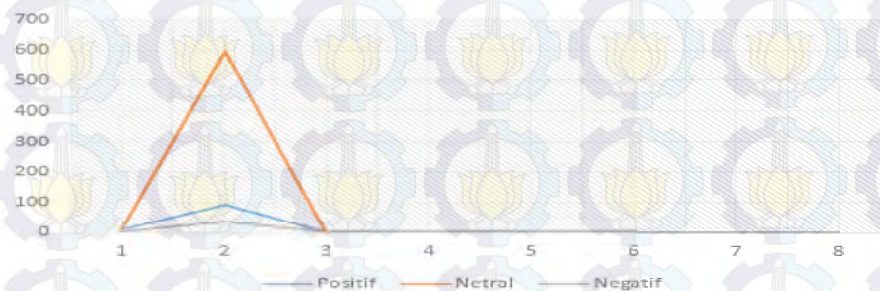
$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.821 * 0.983}{(0.821) + (0.983)} = 0.895$$

Kelas negatif

$$Precision = \frac{TNg}{(TNg+FP+NNg)} = \frac{0}{(0+0+0)} = 0$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{0}{(0+38+0)} = 0$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0 * 0}{(0) + (0)} = 0$$



Gambar 4.1 Percobaan 1

Pada Gambar 4.1 dijelaskan grafik dari hasil percobaan pertama komentar positif yang terprediksi benar berjumlah 8, komentar positif terprediksi netral berjumlah 92 dan komentar netral terprediksi positif berjumlah 10 sedangkan terprediksi netral sebanyak 596 dan komentar negatif sebanyak 38 komentar yang terprediksi netral.

Hasil pengukuran tiap-tiap kelas diperoleh rata-rata *precision*, *recall* dan *F-measure* klasifikasi SVM sebagai berikut :

Data Pelatihan 80% = 744

Data Pengujian 20% = 186

Akurasi klasifikasi = 81.1828%

Rata-rata Precision = 73%

Rata-rata Recall = 81%

Rata-rata F-Measure = 75%

4.3.2 Percobaan Dengan Data Pelatihan 70% dan Data Pengujian 30%

Tabel 4.12 Hasil Percobaan Kedua

Data Aktual	Data Prediksi			Precision	Recall	F-Measure	Kelas
	Positif	Netral	Negatif				
Positif	11	79	0	0.647	0.122	0.206	Positif
Netral	6	522	0	0.823	0.989	0.898	Netral
Negatif	0	33	0	0.000	0.000	0.000	Negatif

Pada Tabel 4.12 dijelaskan percobaan kedua digunakan set data yang telah displit secara *random* diperoleh data pelatihan 70% dan data pengujian 30% yang akan diklasifikasikan menggunakan model pembelajaran SVM maka hasil yang didapatkan berupa *confusion matrix*.

Perhitungan *Precision* hasil berdasarkan Tabel 4.12 dihitung menggunakan persamaan 4.2, 4.3, dan 4.4 dan *recall* menggunakan persamaan 4.6, 4.7 dan 4.8 serta *F-Measure* berdasarkan persamaan 4.9 sebagai berikut :

Kelas positif

$$Precision = \frac{TP}{(TP+NP+FN)} = \frac{11}{(11+6+0)} = 0.647$$

$$Recall = \frac{TP}{(TP+PN+FP)} = \frac{11}{(11+79+0)} = 0.122$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.647 * 0.122}{(0.647) * (0.122)} = 0.206$$

Kelas netral

$$Precision = \frac{TN}{(TN+PN+NgN)} = \frac{522}{(522+79+33)} = 0.823$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{522}{(522+6+0)} = 0.983$$

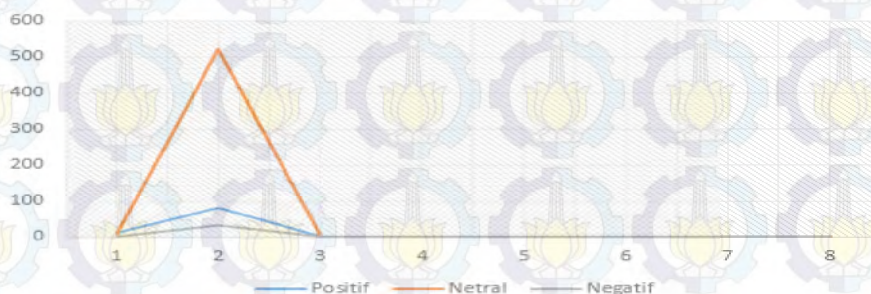
$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.823 * 0.983}{(0.823) * (0.983)} = 0.898$$

Kelas negatif

$$Precision = \frac{TNg}{(TNg+FP+NNg)} = \frac{0}{(0+0+0)} = 0$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{0}{(0+33+0)} = 0$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0 * 0}{(0) * (0)} = 0$$



Gambar 4.2 Percobaan 2

Pada Gambar 4.2 dijelaskan grafik dari hasil percobaan kedua komentar positif yang

terprediksi benar berjumlah 11, komentar positif terprediksi netral berjumlah 79 dan komentar netral terprediksi positif berjumlah 6 sedangkan terprediksi netral sebanyak 522 dan komentar negatif sebanyak 33 komentar yang terprediksi netral.

Hasil pengukuran tiap-tiap kelas diperoleh rata-rata precision, recall dan F-measure klasifikasi SVM sebagai berikut :

Data Pelatihan 70% = 651

Data Pengujian 30% = 279

Akurasi klasifikasi = 81.874%

Rata-rata Precision = 76%

Rata-rata Recall = 82%

Rata-rata F-Measure = 76%

4.3.3 Percobaan Dengan Data Pelatihan 60% dan Data Pengujian 40%

Tabel 4.13 Hasil Percobaan Ketiga

Data Aktual	Data Prediksi			Precision	Recall	F-Measure	Kelas
	Positif	Netral	Negatif				
Positif	12	67	0	0.800	0.152	0.255	Positif
Netral	3	447	0	0.823	0.993	0.900	Netral
Negatif	0	29	0	0.000	0.000	0.000	Negatif

Pada Tabel 4.13 dijelaskan percobaan kedua digunakan set data yang telah displit secara *random* diperoleh data pelatihan 60% dan data pengujian 40% yang akan diklasifikasikan menggunakan model pembelajaran SVM maka hasil yang didapatkan berupa *confusion matrix*.

Perhitungan *precision* hasil berdasarkan Tabel 4.13 dihitung menggunakan persamaan 4.2, 4.3, dan 4.4 dan *recall* menggunakan persamaan 4.6, 4.7 dan 4.8 serta *F-Measure* berdasarkan persamaan 4.9 sebagai berikut :

Kelas positif

$$Precision = \frac{TP}{(TP+NP+FN)} = \frac{12}{(12+3+0)} = 0.800$$

$$Recall = \frac{TP}{(TP+PN+FP)} = \frac{12}{(12+67+0)} = 0.152$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.800 * 0.152}{(0.800) + (0.152)} = 0.255$$

Kelas netral

$$Precision = \frac{TN}{(TN+PN+NgN)} = \frac{447}{(447+67+29)} = 0.823$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{447}{(447+3+0)} = 0.993$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.823 * 0.983}{(0.823) + (0.983)} = 0.898$$

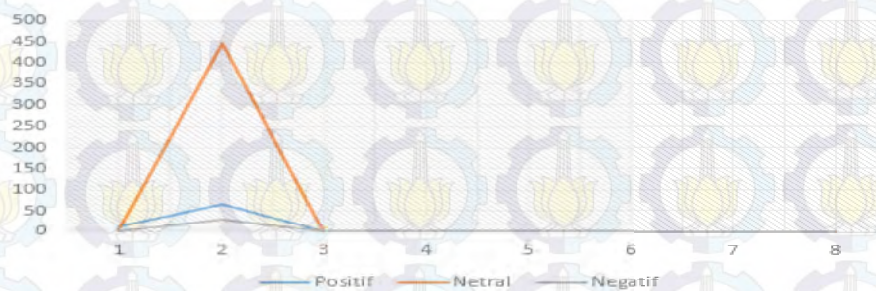
Kelas negatif

$$Precision = \frac{TNg}{(TNg+FP+NNg)} = \frac{0}{(0+0+0)} = 0$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{0}{(0+29+0)} = 0$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0 * 0}{(0) + (0)} = 0$$

Pada Gambar 4.2 dijelaskan grafik dari



Gambar 4.3 Percobaan 3

hasil percobaan ketiga komentar positif yang terprediksi benar berjumlah 12, komentar positif terprediksi netral berjumlah 67 dan komentar netral terprediksi positif berjumlah 3 sedangkan terprediksi netral sebanyak 447 dan komentar negatif sebanyak 29 komentar yang terprediksi netral.

Hasil pengukuran tiap-tiap kelas diperoleh rata-rata *precision*, *recall* dan *F-measure* klasifikasi SVM sebagai berikut :

Data Pelatihan 60% = 558

Data Pengujian 40% = 372

Akurasi klasifikasi = 82.2581%

Rata-rata Precision = 78%

Rata-rata Recall = 82%

Rata-rata F-Measure = 76%

4.3.4 Percobaan Dengan Data Pelatihan 50% dan Data Pengujian 50%

Tabel 4.14 Hasil Percobaan Keempat

Data Aktual	Data Prediksi			Precision	Recall	F-Measure	Kelas
	Positif	Netral	Negatif				
Positif	10	55	0	0.769	0.154	0.256	Positif
Netral	3	373	1	0.827	0.989	0.901	Netral
Negatif	0	23	0	0.000	0.000	0.000	Negatif

Pada Tabel 4.14 dijelaskan hasil percobaan keempat digunakan set data yang telah displit secara *random* diperoleh data pelatihan 50% dan data pengujian 50% yang akan diklasifikasikan menggunakan model pembelajaran SVM maka hasil yang didapatkan berupa *confusion matrix*.

Perhitungan *precision* hasil berdasarkan Tabel 4.14 dihitung menggunakan persamaan 4.2, 4.3, dan 4.4 dan *recall* menggunakan persamaan 4.6, 4.7 dan 4.8 serta *F-Measure* berdasarkan persamaan 4.15 sebagai berikut :

Kelas positif

$$Precision = \frac{TP}{(TP+NP+FN)} = \frac{10}{(10+3+0)} = 0.769$$

$$Recall = \frac{TP}{(TP+PN+FP)} = \frac{10}{(10+55+0)} = 0.154$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.769 * 0.154}{(0.769) * (0.154)} = 0.256$$

Kelas netral

$$Precision = \frac{TN}{(TN+PN+NgN)} = \frac{373}{(373+55+23)} = 0.827$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{373}{(373+3+1)} = 0.989$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.827 * 0.989}{(0.827) * (0.989)} = 0.898$$

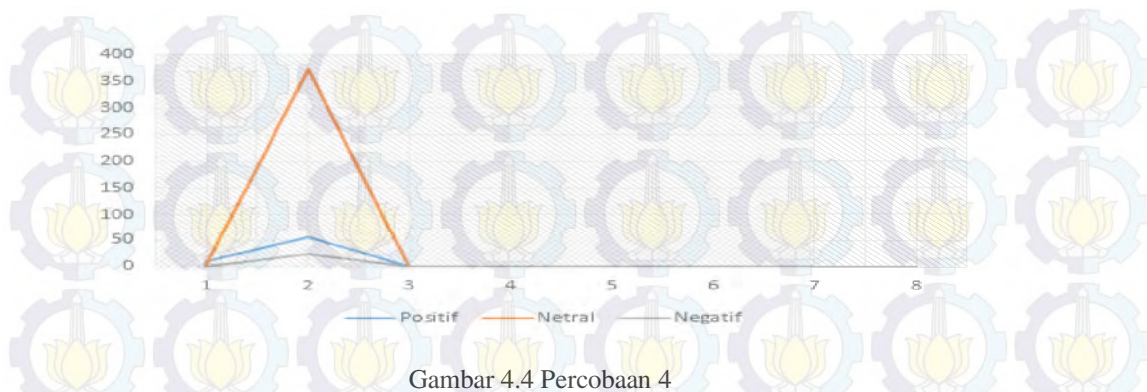
Kelas negatif

$$Precision = \frac{TNg}{(TNg+FP+NNg)} = \frac{0}{(0+1+0)} = 0$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{0}{(0+23+0)} = 0$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0 * 0}{(0) * (0)} = 0$$

Pada Gambar 4.4 dijelaskan grafik dari hasil percobaan keempat komentar positif yang



terprediksi benar berjumlah 10, komentar positif terprediksi netral berjumlah 55 dan komentar netral terprediksi positif berjumlah 3 sedangkan terprediksi netral sebanyak 373 dan komentar negatif sebanyak 33 komentar yang terprediksi netral dan komentar netral yang terprediksi negatif 1.

Hasil pengukuran tiap-tiap kelas diperoleh rata-rata *precision*, *recall* dan *F-measure* klasifikasi SVM sebagai berikut :

Data Pelatihan 50% = 465

Data Pengujian 50% = 465

Akurasi klasifikasi = 82.3656%

Rata-rata Precision = 78%

Rata-rata Recall = 82%

Rata-rata F-Measure = 77%

4.3.5 Percobaan Dengan Data Pelatihan 40% dan Data Pengujian 60%

Tabel 4.15 Hasil Percobaan Kelima

Data Aktual	Data Prediksi			Precision	Recall	F-Measure	Kelas
	Positif	Netral	Negatif				
Positif	8	47	0	0.667	0.145	0.239	Positif
Netral	4	296	1	0.825	0.983	0.897	Netral
Negatif	0	16	0	0.000	0.000	0.000	Negatif

Pada Tabel 4.15 hasil percobaan kelima percobaan kelima digunakan set data yang telah displit secara *random* diperoleh data pelatihan 40% dan data testing 60% yang akan diklasifikasi menggunakan model pembelajaran SVM maka hasil yang didapatkan berupa *confusion matrix*.

Perhitungan *precision* hasil berdasarkan Tabel 4.15 dihitung menggunakan persamaan

4.2, 4.3, dan 4.4 dan *recall* menggunakan persamaan 4.6, 4.7 dan 4.8 serta *F-Measure* berdasarkan persamaan 4.16 sebagai berikut :

Kelas positif

$$Precision = \frac{TP}{(TP+NP+FN)} = \frac{8}{(8+4+0)} = 0.667$$

$$Recall = \frac{TP}{(TP+PN+FP)} = \frac{8}{(8+47+0)} = 0.145$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.667 * 0.145}{(0.667) + (0.145)} = 0.239$$

Kelas netral

$$Precision = \frac{TN}{(TN+PN+NgN)} = \frac{296}{(296+47+16)} = 0.825$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{296}{(296+4+1)} = 0.983$$

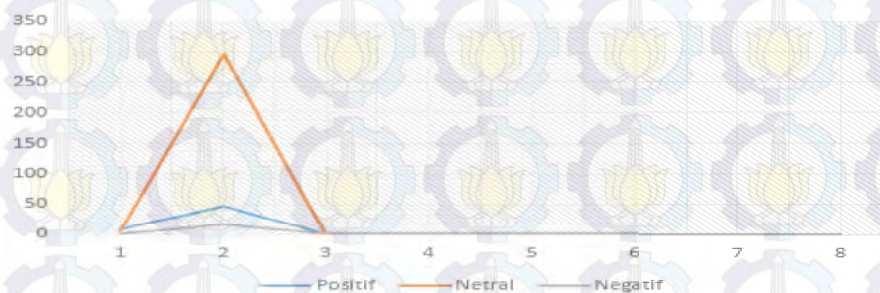
$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.825 * 0.983}{(0.825) + (0.983)} = 0.897$$

Kelas negatif

$$Precision = \frac{TNg}{(TNg+FP+NNg)} = \frac{0}{(0+1+0)} = 0$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{0}{(0+16+0)} = 0$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0 * 0}{(0) + (0)} = 0$$



Gambar 4.5 Percobaan 5

Pada Gambar 4.5 dijelaskan grafik dari hasil percobaan kelima komentar positif yang terprediksi benar berjumlah 8, komentar positif terprediksi netral berjumlah 47 dan komentar netral terprediksi positif berjumlah 4 sedangkan terprediksi netral sebanyak 296 dan komentar negatif sebanyak 16 komentar yang terprediksi netral, komentar netral terprediksi negatif sebanyak 1.

Hasil pengukuran tiap-tiap kelas diperoleh rata-rata *precision*, *recall* dan *F-measure* klasifikasi SVM sebagai berikut :

Data Pelatihan 40% = 465

Data Pengujian 60% = 558

Akurasi klasifikasi = 82.3656%

Rata-rata Precision = 78%

Rata-rata Recall = 82%

Rata-rata F-Measure = 77%

4.3.6 Percobaan Dengan Data Pelatihan 30% dan Data Pengujian 70%

Tabel 4.16 Hasil Percobaan Keenam

Data Aktual	Data Prediksi			Precision	Recall	F-Measure	Kelas
	Positif	Netral	Negatif				
Positif	9	35	0	0.750	0.205	0.321	Positif
Netral	3	221	0	0.828	0.987	0.900	Netral
Negatif	0	11	0	0.000	0.000	0.000	Negatif

Pada Tabel 4.16 dijelaskan hasil percobaan keenam digunakan set data yang telah displit secara *random* diperoleh data pelatihan 30% dan data pengujian 70% yang akan diklasifikasikan menggunakan model pembelajaran SVM maka hasil yang didapatkan berupa *confusion matrix*.

Perhitungan Precision hasil berdasarkan Tabel 4.16 dihitung menggunakan persamaan 4.2, 4.3, dan 4.4 dan recall menggunakan persamaan 4.6, 4.7 dan 4.8 serta F-Measure berdasarkan persamaan 4.17 sebagai berikut :

Kelas positif

$$Precision = \frac{TP}{(TP+NP+FN)} = \frac{9}{(9+3+0)} = 0.750$$

$$Recall = \frac{TP}{(TP+PN+FP)} = \frac{9}{(9+35+0)} = 0.205$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.750 * 0.205}{(0.750) * (0.205)} = 0.321$$

Kelas netral

$$Precision = \frac{TN}{(TN+PN+NgN)} = \frac{221}{(221+35+11)} = 0.828$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{221}{(221+3+0)} = 0.987$$

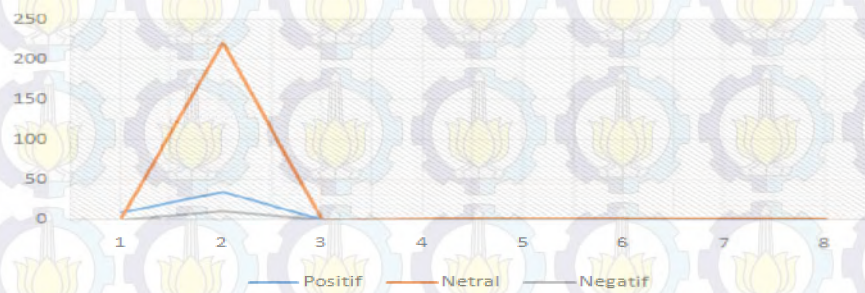
$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.828 * 0.987}{(0.828) * (0.987)} = 0.900$$

Kelas negatif

$$Precision = \frac{TN_g}{(TN_g + FP + NN_g)} = \frac{0}{(0+1+0)} = 0$$

$$Recall = \frac{TN}{(TN + NP + NN_g)} = \frac{0}{(0+11+0)} = 0$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0*0}{(0)+(0)} = 0$$



Gambar 4.6 Percobaan 6

Pada Gambar 4.6 dijelaskan grafik dari hasil percobaan keenam komentar positif yang terprediksi benar berjumlah 9, komentar positif terprediksi netral berjumlah 35 dan komentar netral terprediksi positif berjumlah 5 sedangkan terprediksi netral sebanyak 221 dan komentar negatif sebanyak 11 komentar yang terprediksi netral.

Hasil pengukuran tiap-tiap kelas diperoleh rata-rata *precision*, *recall* dan *F-measure* klasifikasi SVM sebagai berikut :

Data Pelatihan 30% = 279

Data Pengujian 70% = 651

Akurasi klasifikasi = 82.4373%

Rata-rata Precision = 78%

Rata-rata Recall = 82%

Rata-rata F-Measure = 77%

4.3.7 Percobaan Dengan Data Pelatihan 20% dan Data Pengujian 80%

Tabel 4.17 Hasil Percobaan Ketujuh

Data Aktual	Data Prediksi			Precision	Recall	F-Measure	Kelas
	Positif	Netral	Negatif				
Positif	6	21	0	0.857	0.222	0.353	Positif
Netral	1	151	0	0.844	0.993	0.912	Netral
Negatif	0	7	0	0.000	0.000	0.000	Negatif

Pada Tabel 4.17 dijelaskan hasil percobaan ketujuh digunakan set data yang telah displit secara random diperoleh data pelatihan 20% dan data pengujian 80% yang akan diklasifikasikan menggunakan model pembelajaran SVM maka hasil yang didapatkan berupa *confusion matrix*.

Perhitungan Precision hasil berdasarkan Tabel 4.17 dihitung menggunakan persamaan 4.2, 4.3, dan 4.4 dan recall menggunakan persamaan 4.6, 4.7 dan 4.8 serta F-Measure berdasarkan persamaan 4.18 sebagai berikut :

Kelas positif

$$Precision = \frac{TP}{(TP+NP+FN)} = \frac{6}{(6+1+0)} = 0.857$$

$$Recall = \frac{TP}{(TP+PN+FP)} = \frac{6}{(6+21+0)} = 0.222$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.587 * 0.222}{(0.587) * (0.222)} = 0.353$$

Kelas netral

$$Precision = \frac{TN}{(TN+PN+NgN)} = \frac{151}{(151+21+7)} = 0.844$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{151}{(151+1+0)} = 0.993$$

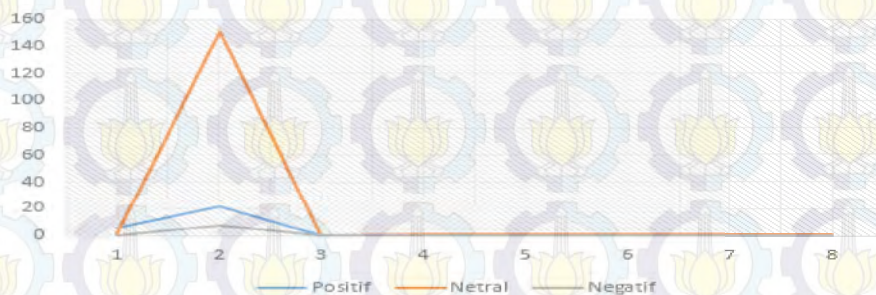
$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0.844 * 0.993}{(0.844) * (0.993)} = 0.912$$

Kelas negatif

$$Precision = \frac{TNg}{(TNg+FP+NNg)} = \frac{0}{(0+1+0)} = 0$$

$$Recall = \frac{TN}{(TN+NP+NNg)} = \frac{0}{(0+7+0)} = 0$$

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision) + (Recall)} = 2 * \frac{0 * 0}{(0) * (0)} = 0$$



Gambar 4.7 Percobaan 7

Pada Gambar 4.5 dijelaskan grafik dari hasil percobaan kelima komentar positif yang terprediksi benar berjumlah 6, komentar positif terprediksi netral berjumlah 21 dan komentar

netral terprediksi positif berjumlah 1 sedangkan terprediksi netral sebanyak 151 dan komentar negatif sebanyak 7 komentar yang terprediksi netral, komentar netral terprediksi negatif

Hasil pengukuran tiap-tiap kelas diperoleh rata-rata *precision*, *recall* dan *F-measure* klasifikasi SVM sebagai berikut :

Data Pelatihan 20% = 186

Data Pengujian 80% = 744

Akurasi klasifikasi = 84.4086%

Rata-rata Precision = 81%

Rata-rata Recall = 84%

Rata-rata F-Measure = 80%

BAB 5

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil akuisisi dan percobaan data teks yang diolah menjadi *sentiment analysis* menggunakan *Support Vector Machine*(SVM) dapat diambil beberapa kesimpulan yaitu :

1. Pengujian menggunakan metode support vector machine (SVM) dilakukan sebanyak 7 kali percobaan dengan tingkat akurasi rata-rata di atas 80%.
2. Dari 7 kali pengujian menggunakan metode *support vector machine*(SVM)akurasi tertinggi 84.4086% dengan data training 20% dan data uji sebanyak 80%.
3. Dalam menentukan nilai dari data Proses *stemming* sangat berpengaruh terhadap hasil karena klasifikasi setelah stemming memiliki akurasi yang lebih rendah dibandingkan dengan data tanpa di stemming.

5.2 Saran

Agar diperoleh hasil yang maksimal, terdapat saran yang dapat digunakan untuk pengembangan lebih lanjut, yaitu :

1. *Sentiment analysis* untuk menghitung frekuensi kemunculan kata sebaiknya dilakukan pengecekan berkali-kali menggunakan kamus untuk pengecekan sinonim kata, dan pengecekan berdasarkan kemiripan makna.
2. Penelitian ini mengklasifikasikan sekumpulan kata-kata atau *bag of word* untuk itu diharapkan pada penelitian selanjutnya memperhitungkan tahapan Pos Tagger yaitu pembagian kelas kata yaitu kata kerja, kata benda, kata sifat dan faktor susunan kalimat yaitu subjek-predikat-objek.
3. Untuk penelitian lebih lanjut diharapkan menggunakan metode pembandingan untuk mengetahui tingkat akurasi setiap metode dan diharapkan dapat membentuk kamus bahasa Indonesia yang dapat menterjemahkan kata-kata saat ini dikarenakan pengguna sering menggunakan kata-kata yang tidak baku sehingga dalam prosesnya dapat mengubah makna kata tersebut
4. Hasil penelitian *sentiment analysis* ini diharapkan menjadi acuan dan suatu kajian studi dengan tujuan menghasilkan suatu kebijakan baru untuk kinerja pemerintahan

LAMPIRAN

Lampiran 1 Hasil Akusisi Data

No	Komentar
1	Air PDAM daerah jagalan - pasar besar - pahlawan, keruh seperti air gotmohon di tindak lanjuti terima kasih
2	Air pdam. Di wilayah tenggumung karya gang 1 dan 2, sudah dua minggu, lancarnya hanya di tengah malam saja, terpaksa harus jaga malam, kurang tidur untuk mendapatkan air pdam
3	Assalamuallaikum....Kapan ya pemerintah kota Surabaya mengadakan penertiban / razia TUKANG PARKIR LIAR.Karena saat ini hampir di setiap tempat selalu di jaga para premanisme berkedok tukang parkir.Dan jujur saja dgn adanya mereka ini
4	Assalamualaikum wr.wbSaya warga dri KEDUNG BARUKRUNGKUT SURABAYAmmeritahukan bhwa Saluran airPDAM di jembatan kedung baruk ygsedang direnovasi saat ini bocor akibatnya air dirumah warga mati lalu smuawarga mengambil air di galianjembatan tersebut ag
5	beberapa hri ini air pam yg keluar keruh dan agak sedikit bau,di rmh saya kedungsroko,dan trkadang tdk keluar air nya.mohon di tindak lanjuti
6	Bu Risma tolong kami...!!!Sungai KARANGMENJANGAN entah kapan terakhir di keruk, yg pasti selama 4 tahun tinggal disini belum sekalipun dilakukan pengerukan. Sedimen sudah tinggi bu, musim penghujan kemarin banjir sekarang harap-harap cemas
7	Bu Risma, kalo mau pasang baru pdam daerah margomulyo, sby, caranya gimana?
8	Bu Risma, mhn bantuan air PDAM, di wilayah RW 08 Kel pegirian tidak lancar, tengah malam baru keluar airnya
9	Jl. Kupang Panjaan gg. 3 kel. Dr. Soetomo Kec. Tegalsari di sepanjang jalan itu sudah banyak lubang dan rusak parah kalau malam gelap jika tidak berhati hati bisa saja jatuh karna tidak ada penerangan.Kalau hujan juga rawan karna banjir
10	Kepada pihak pln atau dkp surabaya, mohon menyeting ulang timer PJU di Jl.Alun-Alun Priok, Perak, karena sampai , saat ini lampu PJU msh menyala padahal langit sdh terang
11	Kepada Yth.Bapak kepala Dinas kebersihan dan Pertamanan Kota Surabaya di Surabaya Dengan Hormat,Mohon bantuan agar Lampu PJU di Jl. Manyar Tirtoyoso Selatan VIII-IX RT 06 RW 05 Kelurahan Manyar Sabrangan Kecamatan Mulyorejo Kota Surabaya (Ko
12	lampu lalu lintas di jln simo persimpangan tol simo, eror. padahal lampu lalu lintas tersebut tergolong baru dipasang dn blum ada 1 bulan beroperasi tp sdh tdk berfungsi lg
13	Maaf mau memberi info kalo PJU sepanjang jalan Sumatra sisi selatan padam hampir 1 minggu ini, sehingga kl malam rawan tindakan kriminal. Mohon bisa ditindaklanjuti dgn dinas terkait. Terimakasih
14	maaf mau nanya, bagaimana cara/prosedur untuk warga miskin mendapatkan BPJS secara gratis (tanpa bayar bulanan).??? .terima kasih atas infonya
15	Selamat pagi, mohon kotoran/sampah hasil kerja bakti pd hari minggu tgl 09/11/2014 segera diambil/dibersihkan, karena 1.) Mengganggu jalan di perempatan Karangrejo gg X dan karangrejo gg VIC, Wonokromo. 2.) Menjadi tempat pembungan sampah orang2 t
16	selamat siang bpk ibu saya mau tanya seputar pasang baru pdam. luas tanah saya 10X10 lebar 10metr sedangkan pipa tersier di depan rumah saya cuma sekitar 4-5 meter dan yang mau saya tanyakan apakah saya sudah bisa pasang sr .sekian terima kasih
17	Selamat sore,kepada YTH PEMKOT,mohon bantuannya karena kami segenap warga sangat merasa terganggu atas aktivitas pengusaha sampah yg kami sampaikan kepada PEMKOT surabaya waktu lalu,kami semua warga berharap agar segera ditindaki karena aktivitas

No	Komentar
18	slmt malam,tolong dong jl gersik / gadukan dekat terowongan tol di kasih lampu penyebrangan atau jembatan penyebrangan , soalnya bahaya sekali bnyk anak sekolah yg membutuhkan itu...apalagi jln itu bnyk truk countainer yg lewat. tolong ya,ini demi keselamatan
19	Sudah hampir satu bulan kami menderita karena tidak lancarnya air PDAM diLingkungan kami (satu Gang),bahkansudah satu minggu ini hampir tidak keluar sama sekal Laporan sudah berulang kali kami sampaikan melalui akun iniatau Pdam Surya Sembada
20	Terima kasih karna masukan saya tentang perbaikan jalan di sepanjang jalan Kupang Panjaan gg. 3 kel. Dr. Soetomo kec. Tegalsari sudah di tangapi.Sudah mulai ada pebaikan dari semalam.Semoga semakin baik. Terima kasih pada instansi terkait
21	@ sapawarga @ suara surabaya tolong diinfokan cara pengurusan SITU atau SKTU di dinas apa ya untuk urusnya?
22	@e100ss @SapawargaSby halte ubhara ada tempat sampah. Senang lihat pemuda bersepeda berhenti sejenak utk buang sampah pic.twitter.com/7YuOYPaMik 12:46 AM - 5 Nov 2014 Details Expand Collapse 0 replies 0 retw
23	"@e100ss @SapawargaSby mhn di bantu pak...di daerah rmh saya listrik blm hidup mulai jam 12 mlm td....di daerah jl gresik,,trima kasih
24	@e100ss @SapawargaSby satpol PP proaktif lah utk tertibkan bendera Golkar di separator jl.urip sumoharjo. Selain memang dilarang jg bahaya,
25	@e100ss @SapawargaSby sptnya pohon2 di sby tdk hanya perlu perantingan tp juga pemangkasan agar tdk menjulang tinggi. Bahaya runkat krn angin,
26	@e100ss @SapawargaSby tahun ini Kalimir belum dikeruk, kt warga Sidodadi X.Biasanya 2-3x /th dikeruk. Gb diambil kmren pic.twitter.com/DLFZ7PvZxq
27	@e100ss Sy urus sendiri ganti KSK, janjinya selesai 3/11, tapi berubah menjd 18/11, apakah mmg sdmkian lama? Kec Sukomanunggl @SapawargaSby
28	@e100ss utk akta online barusan ketemu bag ITnya dispenduk disarankan pakai versi dispendukcapil krn sapawarga lampid msh tahap sosialisasi
29	@indrobaskuworo penutup gorong2 di bawah jembatan KA kertajaya jl.sulawesi membahayakan diharap msyrk pengendara motor hati2 @SapawargaSby
30	"@PDAMSurabaya yang terhormat PDAM kota surabaya, mohon maaf ditempat tinggal saya kenapa airnya susah keluar.bisa keluarnya nunggu jam 12.tq
31	"@roelswafa @SapawargaSby payah pegawai BPN surabaya1,waktu efektif kerja di pakai buat dangdutan,pelayanan kurang maksimal
32	@roelswafa: @SapawargaSby @e100ss lampu penerangan jalan raya sktr TL.jl.kenjeraan arah UNAIR C
33	ada yang mati,Twitter @rolalisasi Sebagian pju jl ngagel jaya arah tl pertigaan ngagel jaya selatan mati. CC : @SapawargaSby
34	"@SapawargaSby – RT @novirizzki: gorong2 yang di darmo satelit arah ke margomulyo tolong cepat diselesaikan dong pagi,sore selalu macet
35	"@SapawargaSby , penyebab terjadinya kemacetan kendaraan di tikungan lampu lalu lintas samping Sungai Kalimas depan Monkasel ??? . Makasih

No	Komentar
36	"@SapawargaSby , saran kepada Dishub untuk melakukan review andal lalin terkait keberadaan pintu masuk parkir sepeda motor Plaza Surabaya
37	"@SapawargaSby ,kepada Ibu Risma,tolong di benahi birokrasi di BPN surabaya1 yg amburadul,apa lg sampai saat ini semua pegawai berjoget ria
38	@SapawargaSby ada dinas terkait masalah tunawisma? Ada nenek nenek tua sakit sering tidur di indomaret sma muhammadiyah 5 (depan mie restu)
39	"@SapawargaSby ada info akun twitter utk polrestabes sby, pln, pdam, yg bersinggungan langsung dg warga. Penting lho supaya cepat terespon
40	@SapawargaSby akhir2 ini kok banyak aduan tentang listrik padam..hari ini wil pandigiling dan kampung malang listrik padam ada apa dgn PLN?
41	@SapawargaSby di per4an Suramadu byk anak2 bersihin mobil pake kemocing maksa minta uang sambil ngintip2 sgt ganggu. Apa itu dibolehkan ya?
42	@SapawargaSby info free hotspot taman sulawesi ada gangguan pengunjung sore jadi agak sepi
43	"@SapawargaSby Jln paving Darmo Permai Utara, bergelombang parah. Itu krn perbaikan pemkot pasca gorong2 tdk sempurna. Tg jwbnya di mana?
44	"@SapawargaSby mati lampu di daerah Kebonsari dan sekitar udah 2 jam lebih,, kira2 kpn nyala nya..??
45	@SapawargaSby mengapa renovasi stasiun semut sangat lama padahal sudah bertahun tahun tidak ada progres untuk segera difungsikan lagi
46	@SapawargaSby mohon informasi dimana mengurus Surat Ijin Tempat Usaha (SITU)
47	"@SapawargaSby mumpung masih jauh sgr bersihkan saluran air, gorong2 & rumah pompa sebelum hujan menghampiri
48	@SapawargaSby sama2...yang di depan jalan mulyorejo no.14 PJU jg sering mati-nyala
49	"@SapawargaSby sdh 1 minggu kualitas air PDAM gak layak pakai,keruh seperti lumpur (daerah pandigiling,kampung malang,kupang)
50	@SapaWargaSBY RT @tweet4nn4: @e100ss jln Tenggilis Mejoyo dpn kampus UBAYA Tenggilis gelap gulita lampu PJU mati sejak seminggu

Lampiran 2 Hasil Filtering Data

aduan	bagus	dinkominfo	lampid	pemotongan
agama	bahagia	dialihkan	mampet	pendemo
agresif	bahaya	digagas	melayani	pembekalan
agung	baiknya	digembosi	membahayakan	pembenahan
ah	bakti	dikerjain	menginspirasi	pemborosan
ajak	balai	diganjar	tersenyum	pembuatan
ajang	balaikota	dikerjain	terselubung	penderitaan
akhirat	bandingin	dikeruk	sapawarga	pendaftaran
akibat	bangkit	dimatikan	sapawargasby	mengapresiasi
akses	bangsa	dinkop	terobosan	menyelenggarakan
akta	banjir	dinsos	merusak	sby
aliran	bantuannya	dirusak	memperjuangkan	sambutannya
amburadul	bekerja	dishub	mengamuk	sandaran
ancaman	bentrok	dishub	membangun	menyiapkan
aneh	berbahagialah	dishub	memuaskan	keterbatasan
anggaran	berbahaya	diskriminatif	mengganggu	kependidikan
angkot	berbakti	dispenduk	pembangunan	kependudukan
angkutan	bergelombang	ditolak	menyelamatkan	menyeret
apresiasi	berharap	diumumkan	menyegel	ketinggalan
berharapan	fanatiknya	kepegawaian	pegawai	ketingian
bermasalah	gangguan	kepemimpinan	menginspirasi	ketrampilan
berpartisipasi	hah	kinerjanya	pemerintahan	memperjuangkan
berjuang	haha	kota	penanaman	memutuskan
bermanfaat	kepeduliannya	kotamadya	mengeluhkan	meminta
berlobang	menyesali	mengeluhkan	mewujudkan	menyeramkan

Daftar Istilah

N_{sv}	: Jumlah Support Vector
α	: Alpha, Pengali Lagrange
i	: Nilai 1,2,3,..., N_{sv}
y	: Label/Kelas dari data
b	: Bias
$(\gamma \cdot x_i^T x + r)^p + b$: Persamaan Polynomial Kernel
TP	: True Positive
FP	: False Positive
TN	: True Negative
FN	: False Negative
w	: Bidang Normal
m	: Jarak antara dua bidang
x_d	: data yang akan diklasifikasikan
x_i, x_j	: vector
ξ_i	: Soft margin hyperplane
df	: Jumlah dokumen dalam term
Lp	: Primal Problem
VC	: Vapnik-Chervonenkis
$R(\alpha)$: Besar kesalahan pada pengujian data(actual risk)
SRM	: Structural Risk Minimization
γ	: Gamma, Parameter
C	: Bounded Support Vector
<i>Training</i>	: Pelatihan
<i>Testing</i>	: Pengujian